View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Saskatchewan's Research Archive

Recognition of Phonemes In a Continuous Speech Stream By Means of PARCOR Parameters In LPC Vocoder

A Thesis Submitted

To the College of Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In the Department of Electrical & Computer Engineering

University of Saskatchewan

Saskatoon, Saskatchewan

By

Ying Cui

© Copyright Ying Cui, January, 2007. All Rights Reserved.

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Electrical & Computer Engineering
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
S7N 5A9

ABSTRACT

Linear Predictive Coding (LPC) has been used to compress and encode speech signals for digital transmission at a low bit rate. The Partial Correlation (PARCOR) parameter associated with LPC that represents a vocal tract model based on a lattice filter structure is considered for speech recognition. For the same purpose, the use of FIR coefficients and the frequency response of AR model were previously investigated.

In this thesis, we investigate the mechanics of the speech production process in human beings and discuss the place and manner of articulation for each of the major phoneme classes of American English. Then we characterize some typical vowel and consonant phonemes by using the eighth order PARCOR parameter associated with LPC.

This thesis explores a method to detect phonemes from a continuous stream of speech. The system being developed slides a time window of 16 ms and calculates PARCOR parameters continuously, feeding them to a phoneme classifier. The phoneme classifier is a supervised classifier that requires training. The training uses TIMIT speech database, which contains the recordings of 630 speakers of 8 major dialects of American English. The training data are grouped into the vowel group including phoneme [ae], [iy] and [uw] and the consonant group including [sh] and [f]. After the training, the decision rule is derived. We design two classifiers in this thesis, one is a vowel classifier and the other one is a consonant classifier, both of them use the maximum likelihood decision rule to classify unknown phonemes.

The results of classification of vowel and consonant in a one-syllable word are shown in the thesis. The correct classification rate is 65.22% for the vowel group. The correct classification rate is 93.51% for the consonant group. The results indicate that PARCOR parameters have the potential capability to characterize the phoneme.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. First of all, I am deeply indebted to my supervisor Dr. Kunio Takaya whose help, stimulating, suggestions and encouragement supported me in all the time of research and writing of this thesis.

Telecommunications Research Laboratories (TRLabs) provided me the financial assistance, the equipment and facilities for the research. I want to thank all the faculty, staff and students in TRlabs for their support to my research. I also would like to thank to the faculty, staff and fellow students in Department of Electrical & Computer engineering.

I am grateful to my husband Quan for his support. Also many thanks go to my parents, who gave me endless support during my study.

Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

AbS	Analysis-by-Synthesis
ADPCM	Adaptive Differential PCM
A/D	Analog to Digital
AR	Autoregressive
ASR	Automatic Speech Recognition
CCITT	International Telegraph and Telephone Consultative Committee
CODEC	Coder-Decoder
CSR	Continuos Speech Recognition
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
IEEE	Institute of Electrical and Electronics Engineers, Inc.
IIR	Infinite Impulse Response
IPA	International Phonetic Alphabet
ITU	International Telecommunication Union
Kbps	Kilobits per second
KHz	kilohertz
LPC	Linear Predictive Coding
MLDR	maximum likelihood decision rule
NIST	National Institute of Standards and Technology
PARCOR	Partial Correlation
PCM	Pulse Code Modulation
RC	Reflection Coefficients
RELP	Residual Excited Linear Prediction/Predictive

SNR	Signal-to-Noise Ratio
SRS	Speech Response System
VOCODER	Voice Coder
ZCR	Zero Crossing Rate

Chapter 1

INTRODUCTION

1.1 LPC Background

The theory of linear predictive coding (LPC), as applied to speech, has been well studied and understood. LPC determines a Finite Impulse Response (FIR) system that predicts a speech sample from the past samples by minimizing the squared error between the actual occurrence and the estimated. The parameters called Partial Correlation (PARCOR) associated with a FIR model represent the basic physical properties, i.e. transmittance and reflectance of the sound wave propagating through the vocal tract. LPC is one of the promising approaches for compressing speech signals and encoding. The LPC encoding is related to analysis of speech whereas decoding corresponds to speech synthesis. [1] The whole system is referred to as a vocoder which is shown in Figure 1.1 The coefficients of the FIR system are encoded and sent. At the receiving end, the inverse system called Autogregressive (AR) model is excited by a random signal to reproduce the encoded speech. In the decoder, excitation and the vocal tract model play important roles to reproduce the speech. The vocal tract is modeled by a time-invariant, all-pole, recursive digital filter over a short time segment (typically 10-30 ms). The time-varying nature of speech is handled by a succession of such filters with different parameters. The excitation is modeled either as a series of pitch pulses (voiced) or as white noise (unvoiced). The use of LPC can be extended to speech recognition since the FIR coefficients are the condensed information of a speech signal of typically 10-30 ms. The Residual Excited Linear Predictive (RELP) vocoder, a class of LPC that uses the residual error signal as a source of excitation, was developed

and reported. In the Residual Excited Linear Predictive (RELP) vocoder, the vocal tract is characterized in the same way as in the pitch-excited LPC. However, instead of switching the source of excitation between pitch pulses for voiced and white noise for unvoiced speech, the residual error signal is used. Since the residual signal becomes a significant amount of data to transmit, RELP is not efficient in compression, but produces more natural speech. [2]

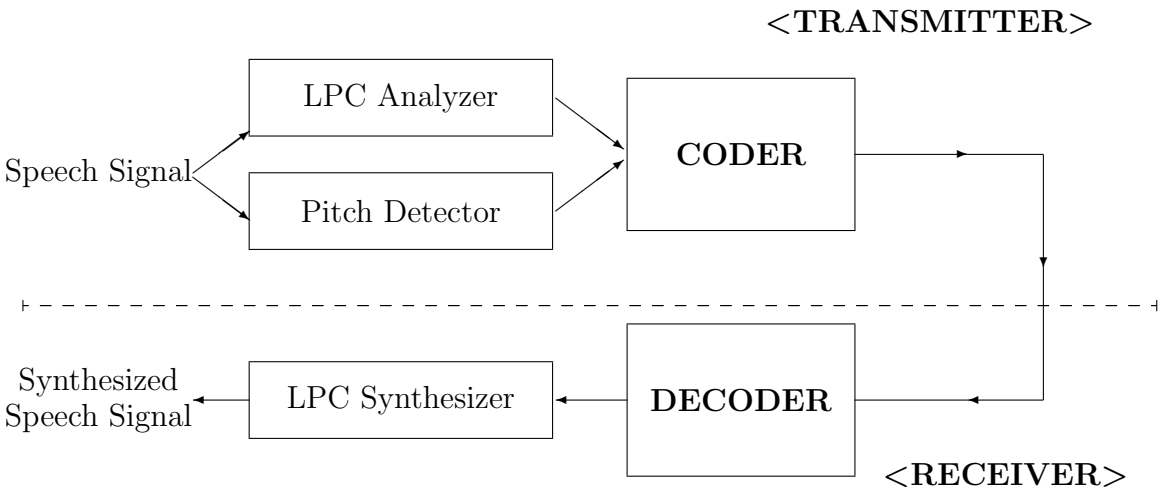


Figure 1.1 An illustration of LPC vocoder

1.2 Speech Recognition

Speech recognition generally may be interpreted as translation of speech signals into linguistic indexes such as words and sentences by machines. Or in other words, it is a speech-to-text conversion problem. The speaker wants his or her voice to be transcribed into text by a machine. It can be used for voice-activated transcription, hearing impaired individuals and telephone assistance. Recently, Research has helped to develop systems that recognize continuous speech in specialized applications, such as to create telephony applications in continuous-speech recognition. For instance, the Uniter Airlines has several speech recognition systems, including employee travel reservations, consumer flight information, up-to-the-minute reporting of lost baggage. Here's an example of a person calling the flight information. When system ask for arrival or departure information, the caller can answer depar-

ture or arrival. When the system ask for the flight number, it's expecting to hear many different responses from callers, such as three sixty-one, three six one and flight three sixty-one etc. Because of this, the speech recognition system should be accountable for "understanding" all possible answers. [3] There are some popular software for speech recognition in the market, such as "Dragon NaturallySpeaking 9" which is developed by InSync Speech Technologies, Inc. It is up to 99% accurate. It is often more accurate and faster than typing because there is no spell mistake and people usually can speak 100 words per minute but type around 40 words. You can use it to dictate letters, e-mails and surfing the web by voice. [4] IBM's ViaVoice is another speech to text software. In order to achieve higher accurate rate, the software need to be trained and listen the target speech in a good environment. People can distinguish a target sound from the interference sound, which is not big enough noise. But for speech recognition by machine, most of the speech recognition system assumes that speech signals have the high signal-noise-ratio (SNR). The high SNR speech signals are extracted and fed into the recognition system or the recognizer. Recently there are some research in recognizing speech sounds in a complex realistic acoustic environments. [5] [6] However, we will not discuss the technologies in this thesis. The ultimate goal is giving computers the ability to act on complex, naturally spoken queries and commands. From the point of system design, the isolated word recognition is more productive and simple. That's why there are are several speech recognition systems available for the variety of the speech recognition task such as the small vocabulary with isolated words or connected words and large vocabulary with isolated words or connected words. In this research, we investigate recognition of phonemes in a continuos speech stream by using PARCOR parameters in LPC vocoder. LPC is not a new technique in speech signal processing, which is used for low-bit rate speech coding and transmission for many years. The speech which is reproduced by LPC synthesize is recognizable, but the rate of understanding synthesis speech signal is around between 70% and 80% in some case.

The speech recognition system involves complicated techniques, which gener-

ally include five modules. The five modules are speech signal processing, feature extraction, segmentation, classification and language model, which are shown in Figure 1.2

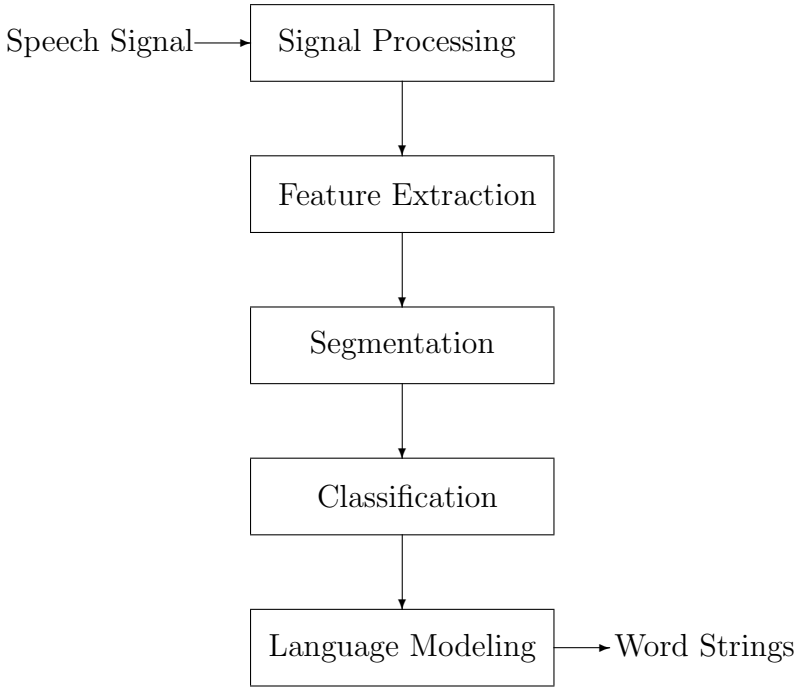


Figure 1.2 A general speech recognition process

The first and second modules, speech signal processing and feature extraction, deal with digitalizing speech signal and processing the sampled speech signal and converting the processed signal into a feature pattern that is suitable for recognition. In general these steps compute a set of parameters, which are the typical representation corresponding to each speech sound. These parameters are often called the features and are generally computed at a short fixed-time interval. In our research, the linear predictive coding (LPC) technique is introduced and the PARCOR parameters are extracted as the features.

In the feature space, the segmentation module partitions the feature pattern into different segments each corresponding to a linguistic unit such as phoneme or word. The classification matches a segment to one of the trained classes such phoneme, words and sentences.

The final language processing stage tries to predict and determine the possible word selections by using the linguistic constraint or rule. [7]

1.2.1 Signal Processing and Feature Extraction

First the A/D (analog to digital) converter is used to digitize the speech signal. The appropriate sample rate must be chosen in order to insure the quality of the speech. The low pass anti-aliasing filter must be implemented before the A/D converter so that the frequencies of the speech signal can be band limited thereby there is not any aliasing between the baseband from $2\pi n/T$ intervals ($1/T$ is the sample rate). [8]

The most sensitive frequency band for the human ear is around 3 KHz. So a 8 KHz sampling rate is enough to provide satisfactory quality speech. A 16 KHz sampling rate provides very high quality speech. When the speech signal is sampled or digitized, we can analyze the discrete-time representation in a short time interval, such as 10-30 ms. Although the speech signal is naturally time variant signal, it can be assumed to be a time invariant signal in short intervals in order to make analysis simple. In a time-varying system, the parameter estimation is fairly difficult. [9]

One of the most important features for speech signal is frequency. A popular method to get the representation of the digitized speech signal in frequency domain is the short time discrete Fourier transform (DFT). The short time spectrum of speech signals can identify the formants, which are considered very important factors to classify the vowels. Formants change as the phoneme class varies, corresponding to the change of the place of articulation, which is mainly determined by the shape of the vocal tract. [10]

The direct digital representation of the short time spectrum can be used as a feature vector for speech recognition, but the feature vector includes too excessive dimensions. A low-dimension feature vector, which can effectively represent the relevant information, is obviously needed. Linear Predictive Coding is one of the effective methods. The LPC based on investigating the human being speech production mechanism provides an efficient parametric model for the physical processes

in the vocal tract. In other words, the vocal tract can be modeled by successive and time-invariant filters, which are characterized by the LPC parameters.

1.2.2 Segmentation and Classification

Segmentation and classification should account for differences in speaker variability, such as pronunciation duration and regional accent differences, in a speaker independent automatic speech recognition (ASR) system.

The segmentation divides the feature pattern into segments or pattern, each segment corresponds to a linguistic unit such as a phoneme or a word. The classification or pattern matching is to match the feature vector pattern into a prescribed class model, which are designed during the training stage. The class model, which usually is represented by a set of parameters, may also be referred to as the template or prototype. There are various ways of classification or pattern matching techniques, which match the unknown pattern into the template or prototype. One basic method in pattern classification is to compare the distance between input pattern and the class model, which is trained prior to classification. In a concrete mathematical way, the distance can be given by the Euclidean distance,

$$d(k) = \sum_{i=1}^{n} [x(i) - c_k(i)]^2 \quad k = 1,\ 2,\ \cdots,\ m \tag{1.1}$$

where $x(i)\ i = 1,\ 2,\ 3\ \ldots,\ n$ is the input feature vector sequence or unknown pattern, $c_k(i)\ i = 1,\ 2,\ 3\ \ldots,\ n\ and\ k = 1,\ 2,\ 3,\ \ldots,\ m$ is the template, which is designed beforehand, the i is the dimension of the feature space and k indicates the different classes.

The classification rule is that if $d(k)$ is the minimum distance, then the unknown pattern x belongs to class k.

1.3 Motivation

For speech recognition, the greatest common denominator of all recognition systems is the signal processing front end, which converts the speech waveform to some type of parametric representation (generally at a considerably lower informa-

tion rate) for further analysis and processing. A wide range of possibilities exists for parametrically representing the speech signal; these include the short time energy, zero crossing rates, short time spectral envelope and other related parameters. In Section 1.1, we mentioned that the FIR coefficients of LPC have the condensed information of a speech signal of typically 10-30 ms. Therefore the LPC is generally considered as the core of the signal processing front end in a speech recognition systems.

LPC provides a good model of the speech signal. The all-pole model of LPC provides a good approximation to the vocal tract spectral envelope. How LPC is applied to the analysis of speech signals leads to a reasonable source-vocal tract separation. As a result, a representation of the vocal tract characteristics becomes possible. The method of LPC is mathematically precise and is simple to implement in either software or hardware. Based on the above considerations, the LPC is used as the signal processing at the front end of recognizers. [10]

1.4 Research Objectives

Investigating human speech production and perception process is very useful to develop a mathematical model, which can realize the recognition. One of the research objectives is to investigate the speech production process of human beings to characterize the acoustic characteristics of some typical vowel and consonant speech sounds by using PARCOR parameters associated with LPC. The speech recognition system is a quite complicated process, particularly the large vocabulary continuous speech recognition (CSR) in automatic speech recognition (ASR). In order to build the complex and big system, we usually start to analyze the related simple and small system. For the speech recognition, we can start to research the recognition of isolated words in small vocabulary. The other objective of this research is to explore a method to classify vowel and consonant phonemes in a one-syllable word in a continuous speech by means of PARCOR parameters.

1.5 Thesis Organization

The thesis is organized into seven chapters.

Chapter 1 introduces the brief background of LPC technique and speech recognition. Motivation and objectives of research are discussed. Thesis organization is given in this chapter.

Chapter 2 discusses the mechanics of speech production process in human beings. The summary of phonemes of American English and the discussion of the place and manner of articulation for each of the major phoneme classes are given in this chapter. Then a simple model for speech production is illustrated.

Chapter 3 introduces Linear Predictive Code (LPC) of speech and its mathematical background. In this chapter, we take a look at liner prediction and the sister topic of autoregressive modeling. In the discussion of linear prediction, an algorithm known as the Levinson recursion is introduced to solve the Normal equations and get the LPC coefficients. Although the original motivation for the Levinson recursion was to provide a fast method to solve the Normal equations, the method brought on other insights are more farreaching. They lead to an efficient lattice structure for the filter associated with PARCOR parameters. Finally, the LPC vocoder model is given.

Chapter 4 describes an experimental method to how to characterize the phonemes by means of the PARCOR parameters. The distributions of PARCOR parameters are presented among different phoneme classes. The experimental results are discussed in this chapter. The potential capability of the PARCOR parameters to characterize phonemes is derived.

Chapter 5 is to explore a method to realize the classification of phonemes in one-syllable word. How to train data and derive the decision rules are discussed. By using the decision rule, which is maximum likelihood decision rule, we design two classifiers, one is vowel classifier and the other one is consonant classifier. After the preprocessing, the test data are fed into the classifiers and the test results are listed and the discussion are presented in this chapter.

Chapter 6 summarizes the research conclusions and the future research directives are suggested.

Chapter 2

PRODUCTION AND BASIC CHARACTERIZATION IN SPEECH SIGNAL

In order to apply digital signal processing (DSP) techniques to the speech signal, it is essential to understand the fundamentals of the speech production process and to find the basic properties of speech sound.

2.1 Speech Production

How do human beings produce speech sounds? In order to answer this question, first we should see the physical and physiological vocal organs. The vocal organs involved in human beings speech production mainly include the lung, larynx, vocal cord and vocal tract. Figure 2.1 illustrates the vocal systems. [11]

Lungs serve as an air reservoir and energy source for the production of speech. The Larynx contains a pair of vocal folds which extend from the thyroid cartilage to the arytenoid cartilages. The space between the vocal folds, called the glottis, is controlled by the arytenoid cartilages. During speech production the vocal folds are on-off to control the vibration and fundamental frequency. The vocal tract can be imagined as a single tube which begins at the vocal folds and ends at the lips with a side branch leading to the nasal cavity. The nasal cavity extends from the velum to the nostrils and assists the vocal tract to produce the nasal sounds of speech. The vocal tract consists of the pharynx which connects the larynx as well as the oesophagus with the mouth or oral cavity. The function of oral cavity is the most important in the vocal tract because its size and shape can be varied by adjusting the relative positions of the palate, the tongue, the lips, the jaws and

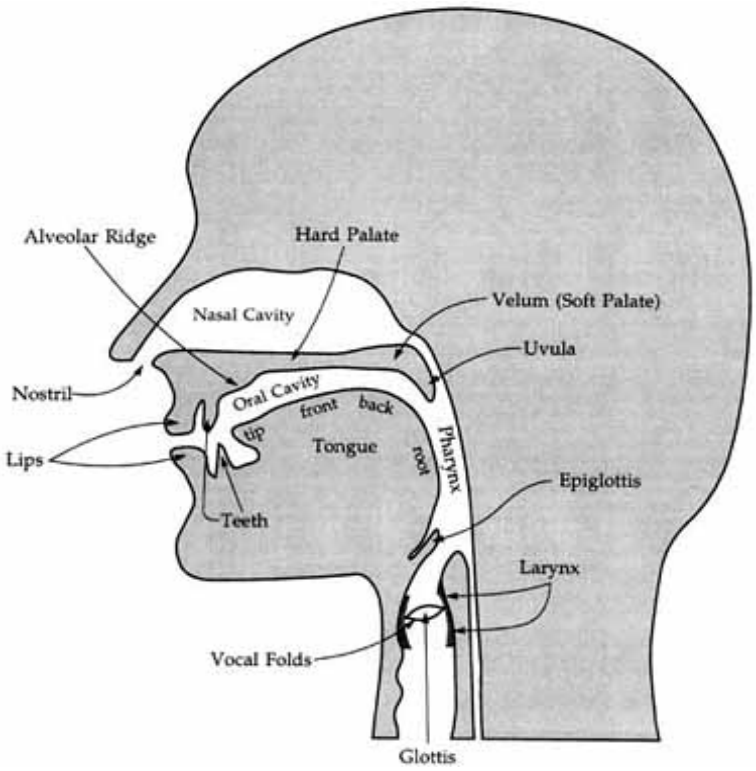


Figure 2.1 The vocal systems of human beings.Source: Department of Linguistics, University of Pennsylvania

the teeth. The length of the vocal tract is typically about 17 centimeters. Speech production is performed during the expiration phase. The expiratory airflow passes through the vocal folds to reach the vocal tract to produce different types of sounds. Speech sounds vary depending on the different manner and place of the human vocal systems, such as vibration vs. no vibration of the vocal folds, front vs. back position of the tongue and stop vs. continuous of the sound. [12]

So far, we briefly discuss how human beings produce speech sounds. There are lots of different languages in the world and each language has its own sounds. In the following, we will discuss the classification of speech sounds in American English.

2.2 Characterization of Speech Sounds

Most languages, including English, can be classified into a set of distinctive sounds or phonemes which are smallest units of speech sounds. One or more phonemes combine to form a syllable, and one or more syllables to combine form a

word. Languages vary in terms of the number of distinct sounds they use. For example, American English has 39 standard phonemes, but Italian has approximately 25 phonemes (depending on the accent). In TIMIT speech corpus of American English, there are 54 basic distinctive sounds. [13] In Table 2.1, the phonetic symbols for American English are listed. [10] In Table 2.1, we use a unique phonetic symbol to represent each distinctive sound. The phonetic symbols are represented in ARPABET. The most common phonetic alphabet is the International Phonetic Alphabet (IPA). Linguists devised the International Phonetic Alphabet (IPA), which is a system of phonetic notation. It is used to accurately and uniquely represent each of the wide variety of sounds used in spoken human language. The IPA is intended as a notational standard for the phonemic and phonetic representation of all spoken languages, but it uses many special characters that are not part of the ASCII character set. So the ARPABET is a widely used phonetic alphabet, which uses only ASCII characters.

There are a variety of ways to classify the speech sounds. From the view of the mode of excitation, we can classify the speech sounds into voiced, unvoiced sounds and plosive sounds. For voiced sounds, the airflow expelled from the lung is forced to pass the glottis with the tension of the vocal folds adjusted so that they generate vibration, thereby producing a quasi-periodic pulse of air as the excitation to the vocal tract. For example, in Table 2.1, the sounds [iy], [ey], and [ae] in the word "bee", "bait" and "cat" are voiced sounds. For unvoiced sounds, the vocal folds are the absence of vibration, and the forcing airflow passes through the constriction which forms at some point in the vocal tract to produce turbulence, also thereby producing a broad-spectrum noise source to excite the vocal tract. In Table 2.1, the sounds labelled [sh], [p] and [f] in the word "shut", "pet" and "fun" are unvoiced sounds. For plosive sounds, by making a complete closure which is usually toward the front of the vocal tract, building up pressure behind the closure and abruptly releasing it, the plosive sounds are produced. The sound [ch] in the word "church" in Table 2.1 is a typical representation.

We can classify sounds into the continuant or the noncontinuous sound. Usually

Table 2.1 Phonetic symbols for American English

Symbol	**Example**	**Symbol**	**Example**
iy	bee	m	mom
ih	bit	n	noon
eh	bet	ng	sing
ae	cat	v	van
aa	bob	dh	that
er	bird	z	zoo
ah	but	zh	azure
ao	bought	f	fun
uw	boot	th	thin
uh	book	s	sat
ow	boat	sh	shut
ay	buy	b	bee
oy	boy	d	dog
aw	down	g	goat
ey	bait	p	pet
w	wit	t	too
l	let	k	kick
r	rent	jh	judge
y	you	ch	church
h	hat		

when producing continuant sounds, the vocal tract keeps a fixed shape excited by the appropriate source. but the noncontinuous sounds are produced by a changing vocal tract shape. [14]

In American English, the phonemes can be classified into the four broad categories: vowels, diphthongs, semivowels, and consonants. In all phonemes of American English, vowels are always voiced and most consonants are unvoiced except some stop and fricative phonemes. Each of the classes can be broken into subclasses according to the manner and place of articulation of the sound within the vocal tract. [10] [1]

2.2.1 Vowels

The principle of vowel production can be described from the excitation as source energy and the place of articulation. Vowels are excited by a quasi-periodic pulse caused by the vibration of the vocal folds. The place of articulation determines the shape of the vocal tract and thereby different sounds are generated by changing the shape of the vocal tract. For vowels, actually the vocal tract shape is relatively fixed and primarily determined by the position of tongue and the positions of the jaw, lips and velum, which also influence the resulting sounds. The resonant frequencies of the vocal tract are decided by the shape of the vocal tract. In the context of speech production, the resonance frequencies of the vocal tract, independent of pitch, are called formants. The pitch and the fundamental frequency (F_0) are often used interchangeably, although there is a subtle difference. Pitch is a perceptual measure, in other words, pitch must be heard and measured by ears connected to a brain. The lowest frequency produced by any particular instrument is known as the fundamental frequency. It does not have to be sensorially perceived. The formants in speech are the resonances in the vocal tract. The summary is that different sounds are produced by the varieties of the vocal tract shape, and the vocal tract shape further determines the formants of the speech sounds, so the formants of the vocal tract are very useful in characterizing each speech sound class and play a very important role in speech recognition. The transfer function of the vocal tract can

determine the spectral envelope of each vowel. When vowels are produced, the vocal tract keeps an essentially fixed shape and the spectra of the vowel are generally well defined, which contributes to the recognition not only for human beings but also for machines.

In terms of tongue position in the oral cavity, vowels are classified into front, central and back three sub categories. Front vowels are [iy], [ih], [eh] and [ae]. The vowels [aa], [er], [ah] and [ao] are mid vowels, and [uw], [uh] and [ow] are back vowels.

2.2.2 Diphthongs

American English has four diphthongs including [ay], [oy], [aw], and [ey] in the respective words "buy", "boy", "down" and "bait", which shown in Table 2.1. The class of the diphthong are transitional sounds. They are produced by starting in a manner and place of articulation of one vowel and ending the articulation position of another vowel. In other words, when diphthong sounds are produced, the vocal tract shape moves smoothly from one vowel to another.

2.2.3 Semivowels

Semivowels lie midway between vowels and consonants. In these phonemes, there is more constriction in the vocal tract than for the vowel, but less than the other consonant categories which will be introduced below. But because of their vowel-like nature, these sounds are called semivowels. They are strongly influenced by the context where they occur which results in a difficulty to characterize. Semivowels consist of the [w] in "wit", the [l] in "like", the [r] in "red", and the [y] in "yes."

2.2.4 Consonants

The principle of consonant production is more complicated than the vowel. We describe it in the following categories: voiced vs. unvoiced; manner of articulation; place of articulation. The place of articulation means where the constriction is located in the vocal tract. The consonants are classified into the following subclasses.

Nasals

Nasals are voiced phonemes which mean the vocal fold vibrate to cause the excitation source air flow. Nasals are generated when the vocal tract is constricted at some point and the velum is lowered (air can flow through the nasal cavity). There are three nasal consonants including [m] in the word "me", [n] in the word "no" and [ng] in the word "sing". The position where the constriction is made in oral cavity for each nasal is different. The constriction of the [m] is at the lips. The constriction of the [n] is just back of the teeth. For the [ng], the constriction is forward of the velum itself.

Fricatives

Fricatives are grouped into two sets, one is the unvoiced fricative and the other one is the voiced fricative.

For unvoiced fricatives, when they are produced, the vocal folds do not vibrate and the vocal tract is excited by a steady air flow which becomes turbulent in the location of the constriction in the vocal tract. The position of the constriction determines which fricative sound is generated. Unvoiced fricatives include [f] in the word "fun", [th] in the word "thin", [s] in the word "set" and [sh] in the word "sheep". The constriction of [f] is located near the lips, for [th] it is near the teeth, for [s] it is close the middle of the oral cavity, and the constriction of the [sh] is located the back of the oral cavity.

For unvoiced fricatives, the broad spectrum noise serves as the source at the position where the constriction is located in the vocal tract. For voiced fricatives, because they are voiced sounds, the excitation source is generated by the vibration of the vocal folds. It makes a significant difference from their unvoiced counterparts. But the place of articulation or the position of the constriction for the two groups fricatives are essentially identical. The counterparts of the unvoiced fricative [f], [th], [s], and [sh] are [v], [dh], [z] and [zh] in the voiced fricatives group. The example words are "vote", "then", "zoo", and "azure".

Stops

There are two subsets of the stop consonants, like the fricative consonants, one set consists of voiced stop consonants, the other one is comprised of the unvoiced stop consonants.

Stop consonants are produced by building up pressure behind some position where a total constriction is located in oral cavity, then abruptly releasing the pressure. Stop consonants are short in duration and are not continuant sounds. Voiced stop consonants include [b], [d], and [g], the corresponding words are "bus", "dog" and "good". For [b] the constriction is at the lips, for [d] it is back of the teeth, and [g] the constriction is close to the velum. The places of constriction of unvoiced stop consonants are similar to voiced stop counterparts. The corresponding unvoiced stop consonants are [p] in the word "park", [t] in the word "ten", and [k] in the word "kite". But the major exception for unvoiced stops is during the pressure builds up and the vocal tract is constricted at some point with the closure of tract, the vocal folds do not vibrate. Even though the vocal tract is closed at some point, the vocal folds are able to vibrate for voiced stop consonants.

Affricates and Whisper

The final two classes of consonants in American English are the affricatives [jh] [ch] and whisper phoneme [h]. The affricate [ch] is unvoiced and dynamical sound. We can model it as a concatenation of the stop [t] and the fricative [sh]. The affricate [jh] is voiced and dynamical sound, too. It can be imaged as the concatenation of the stop [d] and the fricative [zh]. The phoneme [h] is produced without the vocal folds vibrating and by a steady air flow exciting the vocal tract. But the turbulent flow is produced at the glottis. It is not easy to characterize the phoneme [h], since the characteristics of phoneme [h] are similar to those of the vowel which follows phoneme [h]. It means when production of the phoneme [h], the vocal tract assumes the position for the following vowel.

2.3 Vocal Tract Model

We have discussed speech sounds and the way they are produced. We shall consider mathematical models of the process of speech production. In other words, based on the important physical characteristics, realistic and tractable mathematical models should be studied and constructed. Such a model is the basis for the analysis and synthesis of speech. [1] The following block diagram in Figure 2.2 shows the simplified model for speech production.

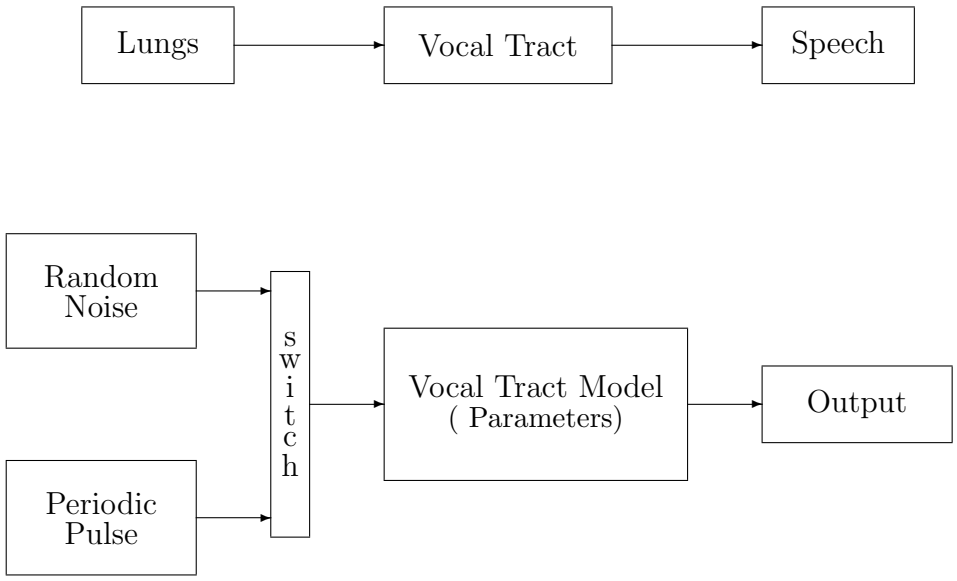


Figure 2.2 Block diagram of the simplified model for speech production

On the top of Figure 2.2, it is a simple block model to represent the speech production process from the physiological view. The corresponding mathematical model is shown at the bottom Figure 2.2. The lungs act as the source of air for exciting the vocal tract. Based on the knowledge that the actual excitation for speech essentially is either a random noise (for unvoiced sounds) or a periodic pulse (for voiced sounds). So we use a switch to chose the excitation source, which is either a random noise or a periodic pulse. In physiologically view, the shape, position and

manner of the vocal tract play an important role to determine the different sounds. In the mathematical way, we need to find a set of parameters to characterize the vocal tract. These parameters can be thought as time-invariant in a short time (10-30 ms).

Chapter 3

LINEAR PREDICTIVE CODING OF SPEECH

3.1 Overview

Before introducing the LPC vocoder, let us to talk about the speech CODEC. The main speech coding techniques are broadly categorized as waveform coding, vocoding and hybrid coding. [15] The idea in waveform coding is signal independent, it attempts to produce a reconstructed signal whose waveform is as close as possible to the original. Waveform codecs have been comprehensively characterized by Jayant and Noll. [16] One of the well known waveform coding is the 64 Kbps PCM (Pulse Code Modulation). It uses non-linear companding characteristics to result in near-constant signal-to-noise ratio (SNR) over the total input dynamic range, which are standardized by the CCITT. The adaptive differential PCM (ADPCM), is standardized by ITU Recommendation G.721. Hybrid coding attempt to fill the gap between waveform and vocoding. The most successful and commonly used are time domain Analysis-by-Synthesis (AbS). [17]

Vocoding uses the knowledge of how the speech signal to be coded was generated, which we discussed in Chapter 2, to extract an appropriate set of source parameters to represent the speech signal to be coded in a given duration of time. In other words, it works in a model form associated with a set of parameters. The vocoder usually is applied to the area of the low bit rate encoding of speech for transmission and storage for computer response systems, for example, the 9.6 Kbps coding by RELP. [2]

The production process of human speech can be modelled in rather detailed mathematical representations, but we need to find out the basic features of the

speech signals in order to further process and analyze. One of the most powerful speech analysis techniques is the method of linear prediction. Linear prediction has been used in numerous problems relating to signal processing. [18] [19] Particularly, in digital processing of speech signals area, the method of linear prediction is used for speech synthesis, recognition, coding and many other applications. [1] [20] LPC determines a FIR model associated with a set of parameters which play a very important role in estimating the basic speech parameters, such as pitch, formants and spectra. The reverse of FIR is called the AR model which is a valid approach to representation of the vocal tract with the excitation.

3.2 Mathematical Background of LPC

3.2.1 Linear Predictive Analysis

Linear prediction estimates the current value of a random sequence $x[n]$ from p previous values of $x[n]$. The estimate $\hat{x}[n]$ can be written as [21]

$$\hat{x}[n] = -a_1 x[n-1] - a_2 x[n-2] - \cdots - a_p x[n-p] \tag{3.1}$$

The prediction error in Equation 3.1 is given by

$$\begin{aligned} \varepsilon[n] &= x[n] - \hat{x}[n] = x[n] + a_1 x[n-1] + a_2 x[n-2] + \cdots + a_p x[n-p] \\ &= \sum_{k=0}^{p} a_k x[n-k] \quad where \quad a_0 \equiv 1 \end{aligned} \tag{3.2}$$

In Equation 3.2, the vector a_k ($k = 0,\ 1,\ 2\ \ldots\ p$) is called linear prediction coefficients. The variance of error $\varepsilon[n]$ is

$$\sigma_\varepsilon^2 = E\{|\varepsilon^2[n]|\} \tag{3.3}$$

The linear prediction parameters consist of linear prediction coefficients and the error variance.

Recall the Equation 3.2, we notice that the linear prediction problem leads to a

FIR filter. The transfer function of the FIR filter is given by [22]

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} \ldots + a_p z^{-p} = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.4}$$

$A(Z)$ is called the prediction error filter. We know that any regular stationary random process can be represented as the output of a linear shift-invariant filter driven by white noise, it is given by [21]

$$x[n] = -\alpha_1 x[n-1] - \alpha_2 x[n-2] - \ldots - \alpha_p x[n-p] + w[n] \tag{3.5}$$

$x[n]$ is called an autoregressive or AR process, the process is " regressed upon itself." It can be seen by comparing Equation 3.2 and Equation 3.5 that if $\alpha_k = a_k$ ($k = 1,\ 2,\ 3,\ \ldots,\ p$), then $w[n] = \varepsilon[n]$. Thus, the transfer function of the AR model which given in Equation 3.5 is an inverse $A(z)$. It can be written

$$H(z) = \frac{1}{A(z)} \tag{3.6}$$

Since $A(z)$ only has the negative powers of z, the AR model is an all-pole IIR filter.

3.2.2 Levinson Recursion

The basic problem of linear prediction analysis we have to solve now is to determine a set of linear prediction coefficients. We use the Orthogonality Theorem to minimize the error variance in order to find the optimal prediction error filter coefficients. The Orthogonality Theorem is give by [21]

$$E\{x[n-k]\varepsilon[n]\} = 0 \quad where \quad k = 1,\ 2,\ \ldots,\ p \tag{3.7}$$

and

$$\sigma_\varepsilon^2 = E\{x[n]\varepsilon[n]\} \tag{3.8}$$

so, we can get the Normal Equations

$$\begin{pmatrix} R_x[0] & R_x[1] & \cdots & R_x[p] \\ R_x[-1] & R_x[0] & \cdots & R_x[p-1] \\ \vdots & \vdots & \vdots & \vdots \\ R_x[-p] & R_x[-p+1] & \cdots & R_x[0] \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.9}$$

where $R_x = E\{xx^T\}$.

The Normal Equations can be solved by the Levinson recursion. The Levinson recursion provides a fast method to solve the Normal equations. It begins with a filter of order 0 and recursively generating filters of order $1, 2, 3$, and so on, up to the desired order p.

The Levinson recursion is introduced in the following description. First Let us consider the forward Normal Equations of order p, which are shown in Equation 3.9. They can be written here as

$$\tilde{\mathbf{R}}_x^{(p)} \mathbf{a}_p = \begin{pmatrix} \sigma_p^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.10}$$

for simplicity, the σ_ε^2 is replaced with σ_p^2. where

$$\tilde{\mathbf{R}}_x^{(p)} = \begin{pmatrix} R_x[0] & R_x[1] & \cdots & R_x[p] \\ R_x[-1] & R_x[0] & \cdots & R_x[p-1] \\ \vdots & \vdots & \vdots & \vdots \\ R_x[-p] & R_x[-p+1] & \cdots & R_x[0] \end{pmatrix} \tag{3.11}$$

and

$$\mathbf{a}_p = \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} 1 \\ a_1^{(p)} \\ \vdots \\ a_p^{(p)} \end{pmatrix} \tag{3.12}$$

Backward prediction predicts the current value by using the "future" points. We

can describe it in a simple mathematical way by using Equation 3.13

$$\hat{x}[n-p] = -b_1 x[n-p+1] - b_2 x[n-p+2] - \cdots - b_p x[n] \tag{3.13}$$

The backward Normal Equations are

$$\begin{pmatrix} R_x[0] & R_x[-1] & \cdots & R_x[-p] \\ R_x[1] & R_x[0] & \cdots & R_x[1-p] \\ \vdots & \vdots & \vdots & \vdots \\ R_x[p] & R_x[p-1] & \cdots & R_x[0] \end{pmatrix} \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \sigma^2_{\varepsilon'} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.14}$$

Similarly, the backward Normal Equations of order p have the form

$$\mathbf{R}_x^{(p)} \mathbf{b}_p = \begin{pmatrix} \sigma'^2_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.15}$$

the $\sigma^2_{\varepsilon'}$ is replaced with σ'^2_p for simplicity, too.

$$\mathbf{R}_x^{(p)} = \begin{pmatrix} R_x[0] & R_x[-1] & \cdots & R_x[-p] \\ R_x[1] & R_x[0] & \cdots & R_x[1-p] \\ \vdots & \vdots & \vdots & \vdots \\ R_x[p] & R_x[p-1] & \cdots & R_x[0] \end{pmatrix} \tag{3.16}$$

and

$$\mathbf{b}_p = \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} 1 \\ b_1^{(p)} \\ \vdots \\ b_p^{(p)} \end{pmatrix} \tag{3.17}$$

Now, we define the term $\mathbf{r_p}$

$$\mathbf{r}_p = \begin{pmatrix} R_x[1] \\ R_x[2] \\ \vdots \\ R_x[p+1] \end{pmatrix} \tag{3.18}$$

Equation 3.11 and Equation 3.16 can be written as

$$\tilde{\mathbf{R}}_x^{(p)} = \left(\begin{array}{ccc} \tilde{\mathbf{R}}_x^{(p-1)} & | & \tilde{\mathbf{r}}_{p-1} \\ & | & \\ - - - - - & | & -- \\ \tilde{\mathbf{r}}_{p-1}^{*T} & | & R_x[0] \end{array}\right) \tag{3.19}$$

and

$$\mathbf{R}_x^{(p)} = \left(\begin{array}{ccc} \mathbf{R}_x^{(p-1)} & | & \tilde{\mathbf{r}}_{p-1}^{*} \\ & | & \\ - - - - - & | & -- \\ \tilde{\mathbf{r}}_{p-1}^{T} & | & R_x[0] \end{array}\right) \tag{3.20}$$

We assume that the linear prediction parameters of order $p-1$ are known. Then think of an augmented set of Normal Equations for the forward problem

$$\tilde{\mathbf{R}}_x^{(p)} \begin{pmatrix} \mathbf{a}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} = \left(\begin{array}{ccc} \tilde{\mathbf{R}}_x^{(p-1)} & | & \tilde{\mathbf{r}}_{p-1} \\ & | & \\ - - - - - & | & -- \\ \tilde{\mathbf{r}}_{p-1}^{*T} & | & R_x[0] \end{array}\right) \begin{pmatrix} \mathbf{a}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma_{p-1}^2 \\ 0 \\ \vdots \\ \triangle_p \end{pmatrix} \tag{3.21}$$

where the

$$\triangle_p = \tilde{\mathbf{r}}_{p-1}^{*T} \mathbf{a}_{p-1} = \mathbf{r}_{p-1}^{*T} \tilde{\mathbf{a}}_{p-1} \tag{3.22}$$

The corresponding augmented set of Normal Equations for the backward linear

prediction problem is given by

$$\mathbf{R}_x^{(p)} \begin{pmatrix} \mathbf{b}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_x^{(p-1)} & | & \tilde{\mathbf{r}}_{p-1}^* \\ & | & \\ - - - - - & | & -- \\ \tilde{\mathbf{r}}_{p-1}^T & | & R_x[0] \end{pmatrix} \begin{pmatrix} \mathbf{b}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} = \begin{pmatrix} \sigma_{p-1}'^2 \\ 0 \\ \vdots \\ \triangle_p' \end{pmatrix} \tag{3.23}$$

where the

$$\triangle_p' = \tilde{\mathbf{r}}_{p-1}^T \mathbf{b}_{p-1} = \mathbf{r}_{p-1}^T \tilde{\mathbf{b}}_{p-1} \tag{3.24}$$

We reverse all of the terms in Equation 3.23 and get this

$$\tilde{\mathbf{R}}_x^{(p)} \begin{pmatrix} 0 \\ -- \\ \\ \tilde{\mathbf{b}}_{p-1} \end{pmatrix} = \begin{pmatrix} \triangle_p' \\ 0 \\ \vdots \\ \sigma_{p-1}'^2 \end{pmatrix} \tag{3.25}$$

We use a constant c_1 to multiply Equation 3.25 and add it to Equation 3.21; the result is

$$\tilde{\mathbf{R}}_x^{(p)} \left(\begin{pmatrix} \mathbf{a}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} + c_1 \begin{pmatrix} 0 \\ -- \\ \\ \tilde{\mathbf{b}}_{p-1} \end{pmatrix} \right) = \begin{pmatrix} \sigma_{p-1}^2 \\ 0 \\ \vdots \\ \triangle_p \end{pmatrix} + c_1 \begin{pmatrix} \triangle_p' \\ 0 \\ \vdots \\ \sigma_{p-1}'^2 \end{pmatrix} \tag{3.26}$$

Now compare Equation 3.26 with Equation 3.10, which is the Normal Equations of order p. Considering the solution to the Normal Equations is unique, then the following results are derived

$$\begin{pmatrix} \sigma_{p-1}^2 \\ 0 \\ \vdots \\ \triangle_p \end{pmatrix} + c_1 \begin{pmatrix} \triangle_p' \\ 0 \\ \vdots \\ \sigma_{p-1}'^2 \end{pmatrix} = \begin{pmatrix} \sigma_p^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.27}$$

and

$$\begin{pmatrix} \mathbf{a}_{p-1} \\ \\ -- \\ \\ 0 \end{pmatrix} + c_1 \begin{pmatrix} 0 \\ -- \\ \\ \\ \tilde{\mathbf{b}}_{p-1} \end{pmatrix} = \mathbf{a}_p \tag{3.28}$$

From Equation 3.27, we can get

$$\sigma_{p-1}^2 + c_1 \triangle_p' = \sigma_p^2 \tag{3.29}$$

and

$$\triangle_p + c_1 \sigma_{p-1}'^2 = 0 \tag{3.30}$$

Similarly, this procedure can be recreated for the backward linear prediction of equations. Reverse Equation 3.21

$$\mathbf{R}_x^{(p)} \begin{pmatrix} 0 \\ -- \\ \\ \\ \tilde{\mathbf{a}}_{p-1} \end{pmatrix} = \begin{pmatrix} \triangle_p \\ 0 \\ \vdots \\ \\ \sigma_{p-1}^2 \end{pmatrix} \tag{3.31}$$

Then this equation is multiplied by a constant c_2 and add it to Equation 3.23,

$$\mathbf{R}_x^{(p)} \left(\begin{pmatrix} \mathbf{b}_{p-1} \\ \\ -- \\ \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ -- \\ \\ \\ \tilde{\mathbf{a}}_{p-1} \end{pmatrix} \right) = \begin{pmatrix} \sigma_{p-1}'^2 \\ 0 \\ \vdots \\ \triangle_p' \end{pmatrix} + c_2 \begin{pmatrix} \triangle_p \\ 0 \\ \vdots \\ \\ \sigma_{p-1}^2 \end{pmatrix} \tag{3.32}$$

In the same way as forward problems, we compare this to the backward Normal Equations 3.15, and the results are given

$$\begin{pmatrix} \sigma_{p-1}'^2 \\ 0 \\ \vdots \\ \triangle_p' \end{pmatrix} + c_1 \begin{pmatrix} \triangle_p \\ 0 \\ \vdots \\ \\ \sigma_{p-1}^2 \end{pmatrix} = \begin{pmatrix} \sigma_p'^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.33}$$

$$\begin{pmatrix} \mathbf{b}_{p-1} \\ \\ -- \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ -- \\ \\ \tilde{\mathbf{a}}_{p-1} \end{pmatrix} = \mathbf{b}_p, \tag{3.34}$$

$$\sigma'^2_{p-1} + c_2 \triangle_p = \sigma'^2_p, \tag{3.35}$$

and

$$\triangle'_p + c_2 \sigma^2_{p-1} = 0. \tag{3.36}$$

To complete the recursion procedure, the consonant c_1 and c_2 are should be found from the Equation 3.30 and Equation 3.36

$$c_1 = -\frac{\triangle_p}{\sigma'^2_{p-1}} \tag{3.37}$$

$$c_2 = -\frac{\triangle'_p}{\sigma^2_{p-1}} \tag{3.38}$$

Because c_1, c_2, $\triangle_p$ and $\triangle'_p$ are defined in terms of the correlation function and the parameters of order $p-1$, these quantities can be computed immediately.

Now let $\gamma_p = -c_1$ and $\gamma'_p = -c_2$, these parameters are known as forward and backward reflection coefficients. The recursion is initialized with Equation 3.39

$$\mathbf{a}_0 = 1; \quad \mathbf{r}_0 = R_x[1]; \quad \sigma^2_0 = R_x[0]. \tag{3.39}$$

The following results can be derived,

$$\gamma_p = \frac{\mathbf{r}^{*T}_{p-1} \tilde{\mathbf{a}}_{p-1}}{\sigma^2_{p-1}} \tag{3.40}$$

$$\mathbf{a}_p = \begin{pmatrix} \mathbf{a}_{p-1} \\ -- \\ 0 \end{pmatrix} - \gamma_p \begin{pmatrix} 0 \\ -- \\ \tilde{\mathbf{a}}^*_{p-1} \end{pmatrix} \tag{3.41}$$

$$\sigma^2_p = (1 - |\gamma_p|^2)\sigma^2_{p-1} \tag{3.42}$$

where the vector $\mathbf{a}_p$ is defined in Equation 3.12. Note that the last element of the vector $\tilde{\mathbf{a}}^*_{p-1}$ is equal to 1, the following result is derived from Equation 3.41,

$$a_p^{(p)} = -\gamma_p. \tag{3.43}$$

From Equation 3.42, since the σ_p^2 and σ_{p-1}^2 are both greater or equal zero, we can draw that

$$|\gamma_p| \leqslant 1 \tag{3.44}$$

Also it implies that

$$\sigma_{p-1}^2 \geqslant \sigma_p^2 \tag{3.45}$$

The recursion for prediction errors can be given by

$$\varepsilon_p[n] = \varepsilon_{p-1}[n] - \gamma_p \varepsilon_{p-1}^b[n-1] \tag{3.46}$$

The γ_p are known as RC (reflection coefficients) because their analogy with similar quantities that occur in the analysis of propagating waves. [1] [23] They are also called partial correlation or PARCOR coefficients because of the statistical interpretation.

3.2.3 Interpretation of the Reflection Coefficients by Partial Correlation

The PARCOR parameter, γ_p, plays an important role in the linear prediction and AR modeling. First, we consider a set of random variables $\{u,\ w_1,\ w_2,\ \cdots,\ w_L,\ v\}$. If u remains correlated with v when the effect of the intermediate variables is removed, this type of correlation is known as *partial correlation*. [21] The following illustration gives us more direct explanations.

As an illustration of this, we suppose there are three random variables u, v, and w and that they have functions like these

$$u = u(w)$$

$$w = w(v)$$

u and v are correlated in general, but if the correlation u with w is removed, then v won't have any influence on u. But if u depends explicitly on both w and v, then we get

$$u = u(w, v)$$

in this case, even if the dependance of u on w is removed, v still has a direct influence on u. This explicit dependence of u on v produces *partial correlation.*

We can develop a geometric picture of partial correlation. It is illustrated in Figure 3.1.

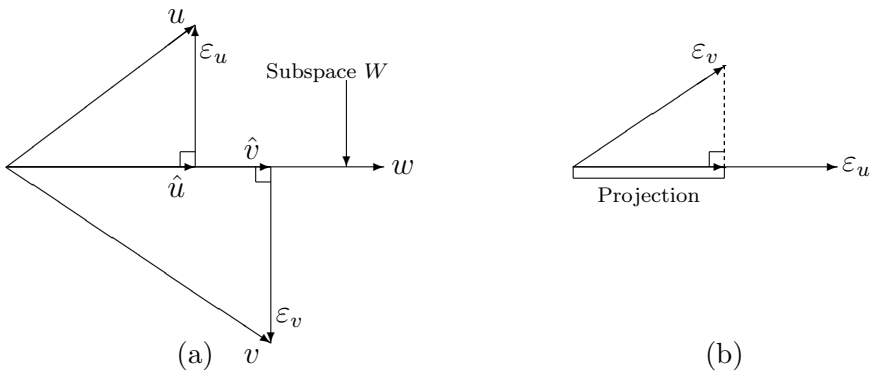


Figure 3.1 Geometric interpretation of partial correlation.(a) Projection of random variables u and v on subspace W and definition of errors.(b) Partial correlation in terms of errors.

In order to remove the influence of the intermediate w_i $(i = 1,\ 2,\ ,\cdots,\ L)$, we project the u and v on the subspace W which is defined by w_i $(i = 1,\ 2,\ ,\cdots,\ L)$. Now, we only deal with the residuals. The estimation errors are given by

$$\begin{aligned} \varepsilon_u &= u - \hat{u} \\ \varepsilon_v &= u - \hat{v} \end{aligned} \tag{3.47}$$

The correlation between the random vectors u and v can be given as

$$E\{uv\} = E\{(\hat{u} + \varepsilon_u)(\hat{v} + \varepsilon_v)\} = E\{\hat{u}\hat{v}\} + E\{\varepsilon_u\varepsilon_v\} \tag{3.48}$$

Because both $\hat{u}$ and $\hat{v}$ lie in the same subspace W and ε_u and ε_v are orthogonal to that subspace, the cross-terms are zero, as can be seen in Figure 3.1(a). On the right of Equation 3.48, the first term represents the indirect correlation because of the presence of the random variables w_i, the second term is the partial correlation. It is the correlation of the errors. Usually, the partial correlation is measured as a normalized quantity which is called the PARCOR coefficient and given by

$$PARCOR[u;v] = \frac{E\{\varepsilon_u \varepsilon_v\}}{E\{|\varepsilon_u|^2\}} \tag{3.49}$$

Recall the definition of the inner product for this vector space, $PARCOR[u;v]$ is the inner product of ε_u and ε_v and normalized by the inner product of ε_u with itself. The magnitude is the ratio of the length of the projection of ε_u on ε_v to the length of ε_u, as can be seen in Figure 3.1(b) The partial correlation is zero when the errors are orthogonal. Also we notice that $|PARCOR[u;v]| \neq |PARCOR[v;u]|$ due to the normalization for general random variables u and v.

Let us consider the $p+1$ data points shown in Figure 3.2 and the associated $(p-1)^{th}$ order forward and backward linear prediction problems. We use u to identify $x[n-p]$ and v to $x[n]$, the points between $x[n-p]$ and $x[n]$ are represented by w_i. In Figure 3.2, the common set C of the intermediate points are used by both the forward prediction of $x[n]$ and backward prediction of $x[n-p]$. The error residuals corresponding to $x[n]$ and $x[n-p]$ are $\varepsilon_{p-1}[n]$ and $\varepsilon'_{p-1}[n-p] = \varepsilon^b_{p-1}[n-1]$. Recall the Equation 3.49, the partial correlation between $x[n]$ and $x[n-p]$ is expressed as

$$PARCOR[x[n-p];x[n]] = \frac{E\{\varepsilon^b_{p-1}[n-1]\varepsilon_{p-1}[n]\}}{E\{|\varepsilon^b_{p-1}[n-1]|^2\}} \tag{3.50}$$

Now we will prove that the quantity in Equation 3.50 is just equal to γ_p. Let us start to look at the full set of points $x[n-p],\ x[n-p+1],\ \cdots,\ x[n]$ in Figure 3.3(a). Note that the backward error $\varepsilon^b_{p-1}[n-1]$ is a linear combination of the points in the set A while the forward error $\varepsilon_p[n]$ is orthogonal to the points in this set. So ,

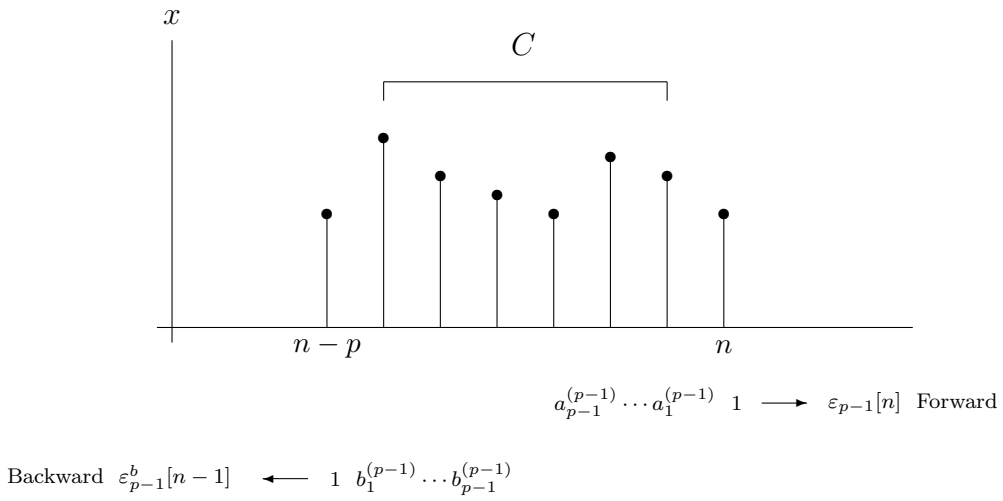


Figure 3.2 Points used for forward and backward linear prediction in interpretation of partial correlation

it shows that

$$E\{\varepsilon_{p-1}^b[n-1]\varepsilon_p[n]\} = 0 \tag{3.51}$$

Recall the Equation 3.46, we substitute it for $\varepsilon_p[n]$ then get

$$E\{\varepsilon_{p-1}^b[n-1](\varepsilon_{p-1}[n] - \gamma_p\varepsilon_{p-1}^b[n-1])\} = 0 \tag{3.52}$$

or

$$\gamma_p = \frac{E\{\varepsilon_{p-1}^b[n-1]\varepsilon_{p-1}[n]\}}{E\{|\varepsilon_{p-1}^b[n-1]|^2\}} \tag{3.53}$$

So far, we prove the result. Now, we will apply the partial correlation to a first-order AR process. In particular, we define the process given by

$$x[n] = \rho x[n-1] + w[n] \tag{3.54}$$

where $w[n]$ is a white noise sequence with mean zero and variance σ_w^2. The corre-

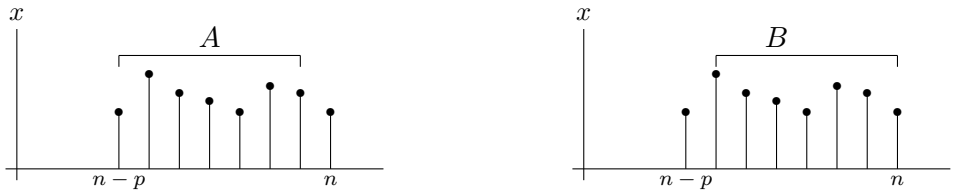


Figure 3.3 Points used in linear prediction

lation function of the resulting random process is

$$R_x[l] = \begin{cases} \frac{\sigma_w^2}{1-|\rho|^2}\rho^l & l \geq 0 \\ \frac{\sigma_w^2}{1-|\rho|^2}\rho^{-l} & l < 0 \end{cases} \tag{3.55}$$

In Figure 3.4, a typical correlation function is illustrated. The $x[n-p]$ and $x[n]$ are correlated for any value of p is obvious and the degree of correlation is represented by the value of the correlation function at $l = p$.

Now recall the partial correlation or γ_p, which is representation of direct influence of $x[n-p]$ on $x[n]$. The coefficients of the first-order AR process are

$$\mathbf{a}_p = \begin{pmatrix} 1 \\ -\rho \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.56}$$

and from Equation 3.43 we get

$$\gamma_p = -a_p^{(p)} = \begin{cases} \rho & p = 1 \\ 0 & p > 1 \end{cases} \tag{3.57}$$

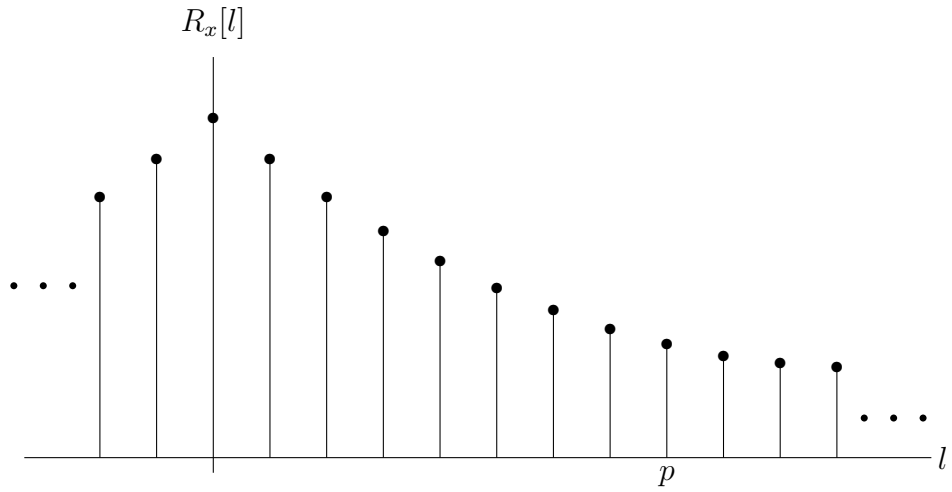


Figure 3.4 Correlation function for a first-order AR process

From Equation 3.57, it shows that the partial correlation of $x[n]$ and $x[n-1]$ is equal to ρ and the partial correlation of $x[n]$ and any earlier points is zero.

3.2.4 Lattice Filter and PARCOR Parameters

We have discussed how to calculate the linear prediction coefficients, once we know these parameters, we can realize the prediction error filter. The direct form of the prediction error filter is shown in Figure 3.5. The corresponding AR model,

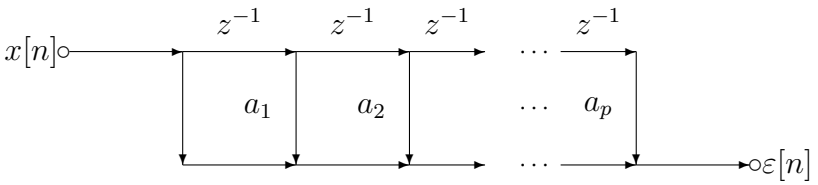


Figure 3.5 Prediction error filter realized by direct form

which is realized in direct form, is illustrated in Figure 3.6. The prediction error filter and the AR model can be realized by the lattice filter structure. The lattice filter is a useful form of a filter representation in digital speech processing. In order to realize both filters by the lattice filter, PARCOR parameters are needed.

In Levinson recursion, the following results can be derived, if we know the $p-1^{th}$ order forward and backward prediction errors, those p^{th} order can be obtained by

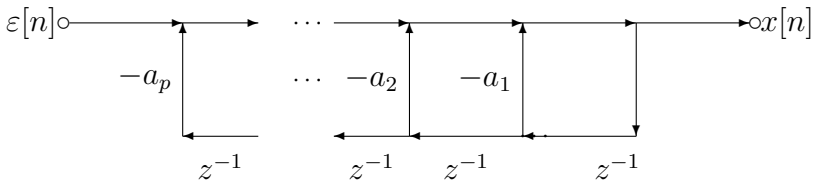


Figure 3.6 AR model realized by direct form

the following equations.

$$\begin{pmatrix} \varepsilon_p[n] \\ \varepsilon_p^b[n] \end{pmatrix} = \begin{pmatrix} 1 & -\gamma_p \\ -\gamma_p & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{p-1}[n] \\ \varepsilon_{p-1}^b[n-1] \end{pmatrix} \tag{3.58}$$

Where $\varepsilon_p^b[n]$ is backward prediction error. Equation 3.58 shows that we can realize the prediction error filter by using the cascading lattice section, which is shown in Figure 3.7. From Equation 3.58, we can get

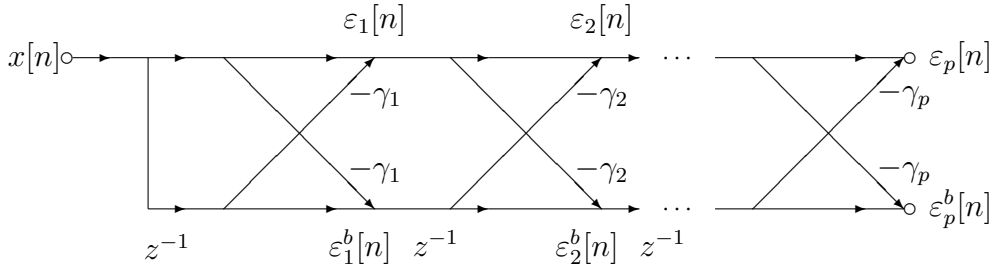


Figure 3.7 Prediction error realized by lattice filter

$$\varepsilon_{p-1}[n] = \varepsilon_p[n] + \gamma_p \varepsilon_{p-1}^b[n-1] \tag{3.59}$$

The AR model also can be realized in lattice form by inverting the structure of Figure 3.7. The AR model realized by lattice filter is shown in Figure 3.8.

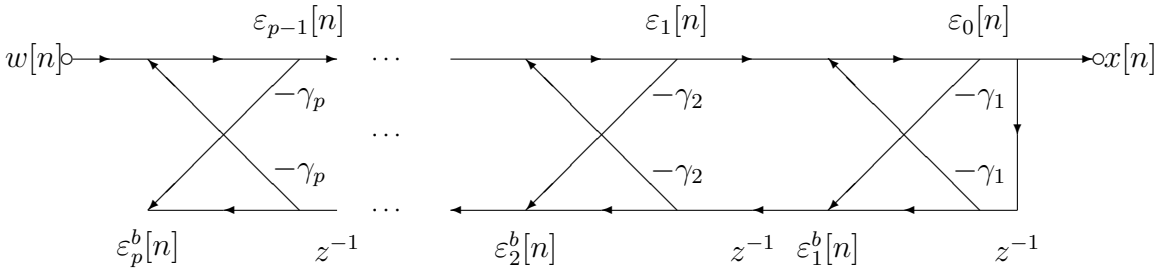


Figure 3.8 AR model realized by lattice filter

3.3 LPC Vocoder

First, a typical LPC vocoder is illustrated in Figure 3.9. The LPC analyzer can be detailed in Figure 3.10.

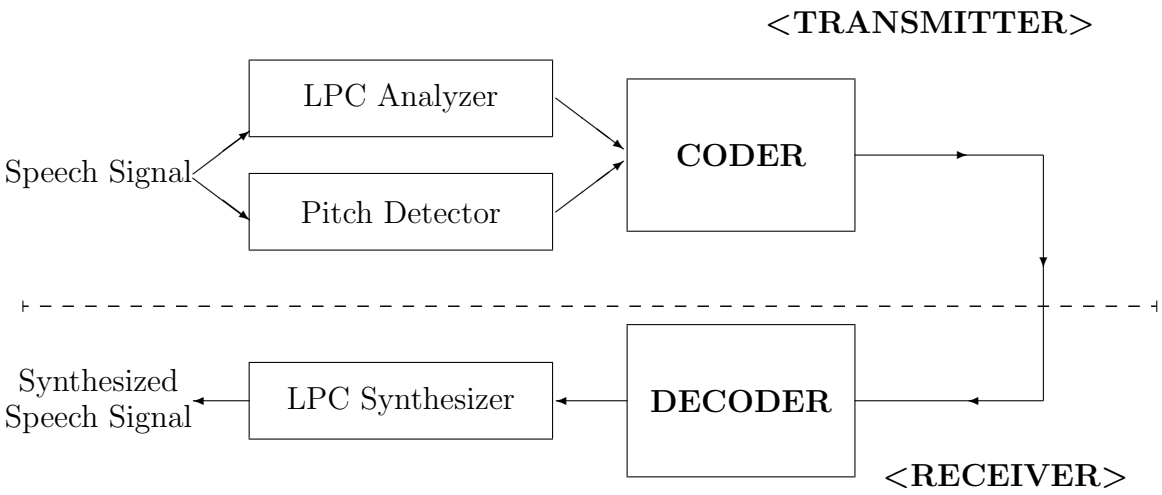


Figure 3.9 LPC vocoder block diagram

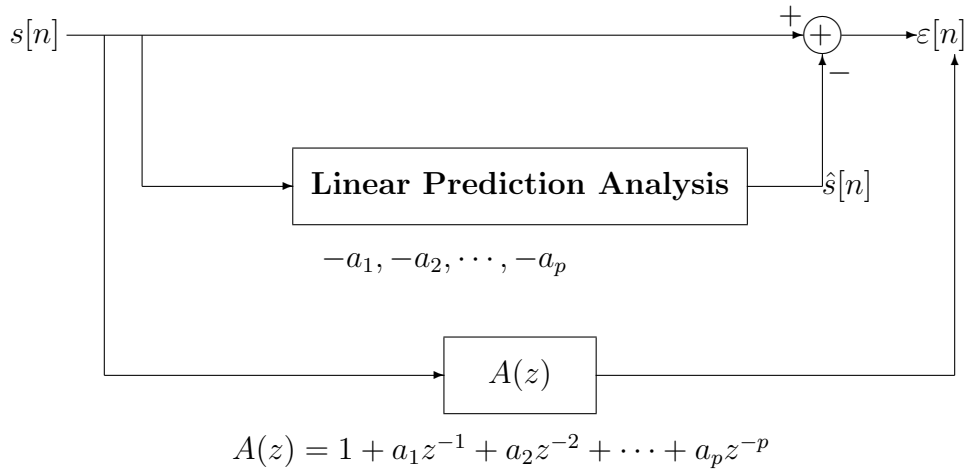


Figure 3.10 LPC analyzer

In Figure 3.9, the vocoder consists of two parts, the transmitter and the receiver. The transmitter performs LPC analysis and pitch detection, then codes the parameters for transmission. The choice of the order p in LPC analyzer is an important consideration. If the order p is in the range of 8 to 10, the input speech signal

can be represented well by the LPC parameters. [1] The prediction error is a good approximation to the excitation source in the receiver of LPC vocoder. The prediction error signal is expected that to be large (for voiced sounds) at the beginning of each pitch period. By detecting the positions of the samples of prediction error which are high value, we can determine the pitch period.

The receiver decodes the parameters and synthesizes the output speech from them. In the receiver, the excitation source, which is either a white noise (for unvoiced sounds) or a periodic pulse (for voiced sounds), goes through the LPC synthesizer. LPC synthesizer is the inverse of $A(z)$. It is called AR model, and its transfer function is $H(z) = A(z)^{-1}$. In order to produce the speech-like signal, the excitation and the AR model have to vary with time since the speech signal is the time-varying signal in nature. But it is reasonable to assume that the general properties of the excitation and vocal tract remain fixed for a short time, such as 10 to 30 ms. So a time-invariant AR model excited by an excitation signal which switches from quasi-periodic pulse for voiced speech to random noise for unvoiced speech is used to model the speech production in a short time. The synthesized speech signal is produced at the output of the AR model.

LPC vocoder is well applied in the low bit rate transmission and speech response system (SRS). Since we have known that the vocal tract imposed its resonances on the excitation to produce different sounds by varying the shape of vocal tract, the poles of transfer function in AR model correspond to the resonances (formants) of speech sound, we can consider applying these parameters' information to characterize speech sound in speech recognition systems. The LPC parameters have the ability to characterize the speech signals in speech recognition systems, particularly for vowels in the phoneme level. [24] In the LPC parameters, one of the important parameters is the PARCOR parameter associated with the AR model. It is the representation of speech physical characteristics. [25] [26]

Chapter 4

IMPLEMENTATION OF PARCOR IN SPEECH SIGNALS

In Chapter 2, we discuss the acoustic characterization of various phoneme classes from the manner and place of articulation. In other words, it is observed from the sound production process of the human being in real life. From the mathematical view, the speech signal can be represented in some type of parametric form for further analyzing and processing. There are a wide range of possibilities for representing the speech signals in the mathematical way, one of them is short time spectral envelope in speech signal processing and analyzing. Linear predictive coding (LPC) is another important and dominant technique in analyzing and processing speech signals, which is introduced in Chapter 3.

In this chapter, we explore a method to illustrate the distributions of PARCOR parameters of some typical phonemes obtained by the LPC technique. The PARCOR parameters are calculated by the auto correlation method which was discussed in 3. The corresponding waveforms and short time spectral characterizations of the typical phonemes are illustrated, too. Then we summarize the characteristics of different speech sound in the phoneme level by analyzing the parametric representation. Finally, we present correlations among eight PARCOR parameters in a two-dimensional space.

4.1 Acoustic-Phonetic Characterization

In the experiments, the speech signals were chosen from a continuous stream of speech in the TIMIT database. TIMIT contains total of 6300 sentences, 10 sentences

spoken by 630 speakers from 8 major dialect regions of the United States. [13] The speakers of dialect region distribution and phonemic and phonetic symbols are listed in Appendix A. The speech in TIMIT database is sampled at a 16K sampling rate. First, we extracted single phonemes and categorized them into two groups. One is the vowel group, and the other one is the consonant group. There are vowel phonemes [ae], [iy] and [uw] in the vowel group. The consonant group includes fricative consonants [sh] and [f].

Each phoneme sound in both groups was spoken by a female and a male speaker. The speakers are from different dialect regions. Then we segmented each single phoneme utterance into consecutive frames and each frame has 256 samples. Since the speech sampling rate is 16K in TIMIT database, the 256 samples frame is 16 ms in duration. We mentioned that if limited to a 10-30 ms short time, the speech signal can be characterized as a time-invariant signal, the 16 ms duration falls into that range. For each utterance, the number of frames is different. This is because the sounds are different in duration. Even for the same phoneme sound, different speakers produce slightly different duration. Also it is very natural that the duration varies from time to time when the same person produces the same phoneme repeatedly. In Figure 4.1 to Figure 4.10, we show the waveform, FFT spectra and the eighth-order PARCOR parameters distributions for each frame. In Table 4.1, we list the information of data such as the speaker name, the dialect region, the gender, and the word used to extract the phoneme in Figure 4.1 to Figure 4.10. In each figure, the number of sub-figures is different, but in each sub-figure, at the top rows are the waveform plots, which are normalized amplitude signal between -1 and $+1$. The spectra of the corresponding waveforms are illustrated in the middle rows, which are square of frequency response of normalized signal shown in top rows. At the bottom rows are the eighth-order PARCOR parameters distributions associated with the LPC technique. From Equation 3.44, we know that the PARCOR parameters are between -1 and $+1$.

The consecutive frames of 256 samples, which are generated by segmenting each single phoneme sound, are illustrated in numeric order on the top of the sub-figures.

Table 4.1 Information of data in Figure 4.1 to Figure 4.10

Figure Label	Phoneme	Region	Speaker Name	Gender	Extract Word
Figure 4.1	[ae]	DR5	fkkh0	Female	Cat
Figure 4.2	[ae]	DR4	mcss0	Male	Cat
Figure 4.3	[iy]	DR3	falk0	Female	Greasy
Figure 4.4	[iy]	DR4	mbma0	Male	She
Figure 4.5	[uw]	DR2	flma0	Female	Moon
Figure 4.6	[uw]	DR6	mrxb0	Male	Moon
Figure 4.7	[sh]	DR4	falr0	Female	She
Figure 4.8	[sh]	DR2	mcew0	Male	She
Figure 4.9	[f]	DR4	falr0	Female	Enough
Figure 4.10	[f]	DR7	mbbr0	Male	Enough

For the same phoneme, there are two sets of figures to show the characterization of the waveform, corresponding spectra and PARCOR distributions because it is spoken by two different speakers.

4.1.1 Vowels

Vowel [ae]

In Figure 4.1, the vowel [ae], which is extracted from the word "cat", is spoken by a female speaker from dialect region five. There are 9 consecutive frames in Figure 4.1, it means this vowel sound lasts around 144 ms (16 ms/frame* 9 frames) in duration. Given that the waveforms, spectral shape and PARCOR parameters distributions in nine frames are similar to each other, for simplicity, we pay attention to frame 3 in Figure 4.1 (a). It can be observed that the waveform is periodic, the spectral shape is well defined. The first to the fourth PARCOR parameters alternately distribute between positive and negative. The first PARCOR parameter is close to -1. The second PARCOR parameter is close to $+0.9$. The third is located around -0.4 and the fourth goes up to $+0.5$. The fifth, seventh and eighth are all around zero. The sixth is located around $+0.4$.

The waveforms, spectra and PARCOR distributions of another vowel [ae] is illustrated in Figure 4.2. This vowel [ae] is extracted from word "cat", too, but it is spoken by a male speaker from dialect region four. Also, there are 9 consecutive

frames in Figure 4.2, so it lasts the same 144 ms in duration. We pick frame 3 in Figure 4.2 (a) to analyze; similarly, the periodic characteristics are shown in the waveform plot. The first four PARCOR parameters distribute alternately between negative and positive. The fifth is close to zero. But the distributions of the sixth, the seventh and the eighth PARCOR parameters are slightly different between Figure 4.1 and Figure 4.2, it is natural that the same sound spoken by different people produces difference.

Vowel [iy]

In Figure 4.3 and Figure 4.4, waveforms, spectra and PARCOR distributions of vowel [iy] are illustrated. The vowel [iy] in Figure 4.3 is extracted from the word "greasy" and is spoken by a female from dialect region three. In Figure 4.4, the vowel [iy] is extracted from the word "she" and is spoken by a male from dialect region four. In both figures, there are two sub-figures and 6 consecutive frames, the duration of the two vowel [iy] is same, and it is 96 ms. The waveform plots in both figures show periodic characteristics. We draw our attention to PARCOR distributions in Figure 4.3, the first, the third and the fifth PARCOR parameters in first five frames are all negative and located between -0.5 and -0.7. Only the fifth PARCOR parameter of the frame 6 in Figure 4.3 is located at -0.2. Except for frame 6, the second PARCOR parameters in all frames are all positive and are located around $+0.4$, the second PARCOR in frame 6 is a little bit lower than others, it is close to $+0.2$. From frame 1 to frame 6 in Figure 4.3, the seventh and eighth PARCOR are all positive values. The sixth PARCOR parameter in frame 1 is around zero, in frame 2 is about $+0.3$, in other frames, they are all negative between -0.3 and -0.2. Although the PARCOR distributions in all frames are slightly different, they are similar to each other in general. Now we turn to the distributions of PARCOR parameters in Figure 4.4, all of the first PARCOR parameters are close to -1 , all of the second PARCOR parameters are around $+0.4$ and all of the third PARCOR parameters are around -0.5 in all six frames. The fourth PARCOR parameters in frame 1 and frame 2 are close to zero, but in frame 3 to frame 6 are

all around +0.2. The fifth PARCOR parameters are negative and located between −0.1 and −0.3 from frame 2 to frame 6. In frame 1, the fifth PARCOR parameter is around zero. The sixth PARCOR parameters in all frames are all negative and located between −0.3 and −0.1. The seventh PARCOR parameters in all frames excluding the frame 2 are all positive and close to +0.2. All of the eighth PARCOR parameters are located around +0.4.

Vowel [uw]

In Figure 4.5 and Figure 4.6, waveforms, spectra and PARCOR distributions of the vowel [uw] are illustrated. The vowel [uw] in Figure 4.5 is extracted from the word "moon" and is spoken by a female from dialect region two. In Figure 4.6, the vowel [uw] is extracted from the word "moon", too, but it is spoken by a male from dialect region six. In Figure 4.5, there are two sub-figures and 6 consecutive frames, the duration of the vowel [uw] is 96 ms. In Figure 4.6, there are four sub-figures and 12 consecutive frames, the duration of the vowel [uw] is 192 ms. The waveform plots in both figures show the periodic characteristics. When we observe the eighth PARCOR parameter distributions in Figure 4.5, the first to the fourth PARCOR parameters in all frames distribute similarly. All of the first PARCOR parameters are close to −1. For the second parameters in all frames, they are all located about +0.5. The third parameter in each frame is around −0.15 and the fourth is around +0.4 in each frame. The sixth parameters are located −0.3 in frame 1, frame 2, frame 3 and frame 6, but in frame 4 and frame 5, they are very close to −0.1. All of the seventh and the eighth PARCOR parameters swing between +0.1 and −0.1 in Figure 4.5. In Figure 4.6, the PARCOR distributions are similar to those in Figure 4.5. All of the different PARCOR parameters are close to −1. For the second parameters in all frames, they are all located about +0.5. The third parameter in each frame is between −0.1 and +0.1. The fourth and the fifth are all positive and located around +0.3. All the sixth, the seventh and the eighth PARCOR parameters fluctuate between −0.1 and +0.1.

By comparing the PARCOR distributions of vowel [ae], [iy] and [uw] in Figure

Figure 4.1 to Figure 4.6, we notice that the PARCOR distributions of different vowels are quite different, but for the same vowel such as vowel [uw] in Figure 4.5 and Figure 4.6, they are only slightly different. The distributions of PARCOR parameters have the potential ability to distinguish the vowel [ae], [iy] and [uw].

4.1.2 Consonants

From Figure 4.7 to Figure 4.10, waveforms, spectra and PARCOR distributions of consonant [sh] and [f] are shown.

Consonant [sh]

In Figure 4.7, a consonant [sh] is extracted from word "she" and spoken by a female from dialect region four. Another consonant [sh] in Figure 4.8 is extracted from word "she", too, but it is spoken by a male speaker from dialect region two. There are 6 and 4 consecutive frames in Figure 4.7 and Figure 4.8 respectively, so it lasts 96 ms and 64 ms in duration respectively. The non-periodic nature is obvious in the waveform plots in both figures. We can see the broad-band noise spectra in the middle of row in each frame. In Figure 4.7, except the seventh parameter in frame 1, all of the PARCOR parameters distribute in the positive space. All of the first parameters in all frames are about $+0.2$, the second are around $+0.7$. In Figure 4.8, all of the first PARCOR parameters are located -0.4, the second to the seventh in each frame have the similar distributions as those in Figure 4.7.

Consonant [f]

The consonant [f] is extracted from the word "enough" and spoken by a female from dialect four is shown in Figure 4.9. Another consonant [f], shown in Figure 4.10, is extracted from word "enough", too, but it is spoken by a male speaker from dialect region seven. There are 6 and 5 consecutive frames in Figure 4.9 and Figure 4.10 respectively, so it lasts 96 ms and 80 ms in duration respectively. In the waveform plot of each frame, the non-periodic nature is noticeable in both figures. Also we can see the broad-band noise spectra in the middle of the row in each frame. In Figure 4.9, except for the first parameters in frame 1 and frame 2, all of the PARCOR parameters fluctuate around zero. Turn to Figure 4.10, we see the

similar situation, except the first PARCOR parameters in frame 1 and 2, all of the PARCOR parameters swing around zero.

4.1.3 Vowels and Consonants

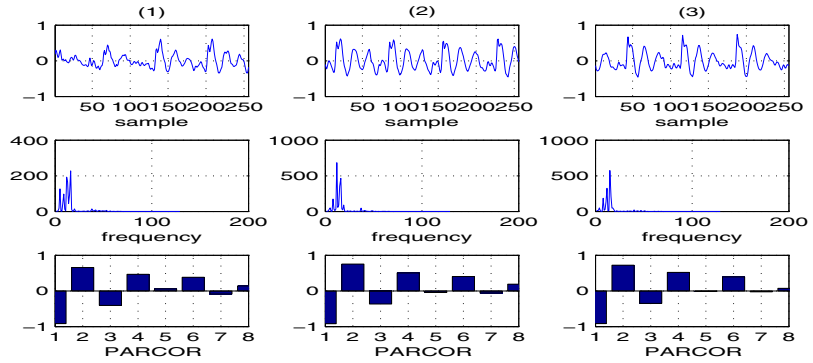
By observing the duration in Figure 4.1 to Figure 4.10, generally speaking, the vowel [ae], [iy] and [uw] is longer than consonant [sh] and [f] in duration. In Figure 4.1 and Figure 4.2, there are nine consecutive frames , which means both vowels last around 144 ms. We look at the vowel [iy] in Figure 4.3 and Figure 4.4, they both have six consecutive frames and last 96 ms . In Figure 4.5 and Figure 4.6, there are six and twelve consecutive frames , so the vowel [uw] last 96 ms and 192 ms respectively. For the consonant [sh] and [f], in Figure 4.7 and Figure 4.9, there are six consecutive frames and both of them last 96 ms in duration, but in Figure 4.8 and Figure 4.10 , there are only four and five consecutive frames, it means they last 64 ms and 80 ms respectively.

If we compare the waveforms in vowel figures (from Figure 4.1 to Figure 4.6) with waveforms in the consonant figures (from Figure 4.7 to Figure 4.10), the periodic characteristics are obvious in vowels, but for consonants [sh] and [f], the non-periodic nature is noticeable. The spectra of vowel [ae], [iy] and [uw] are shown in the middle of each sub-figure from Figure 4.1 to Figure 4.6 are well defined. This is because that the vowels are generated by exciting an essentially fixed vocal tract shape with the quasi-periodic pulsed of air caused by the vibration of the vocal folds. But for the consonants, there are broad band noise spectra in the middle of each sub-figure from Figure 4.7 to Figure 4.10. These consonants are generated by exciting the vocal tract with a steady air flow, which becomes turbulent in the location of the constriction in the vocal cavity. That's why we can see the broad-band noise spectra in the middle of each sub-figure.

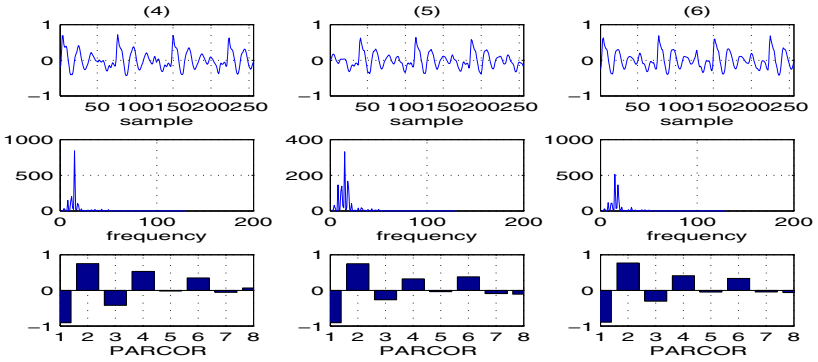
The distributions of PARCOR parameters of vowel [ae], [iy] and [uw] in Figure 4.1 to Figure 4.6 are quite different from consonant [sh] and [f] in Figure 4.7 to Figure 4.10. Generally, the PARCOR parameters (specially, the first four PARCOR parameters) in vowel [ae], [iy] and [uw] alternately distribute between -1 and $+1$,

but for the consonant [sh] and [f], most of the eight PARCOR parameters distribute close to zero.

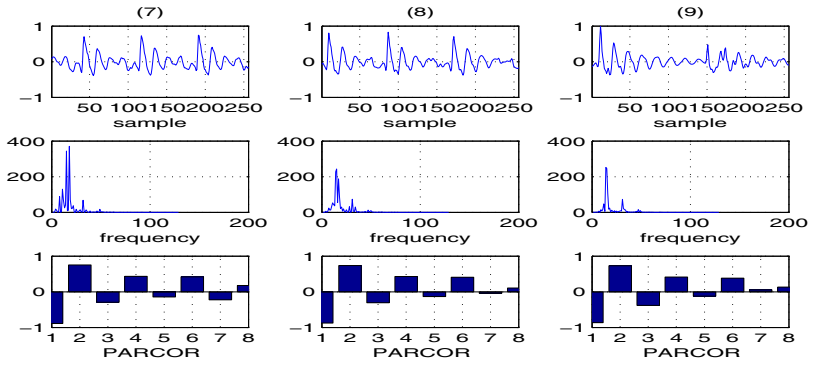
So far, we can tell the difference of the eighth-order PARCOR distributions by observing each phoneme in all figures. But it is difficult to differentiate them by observing them in Figure 4.1 through Figure 4.10, since only two samples in each phoneme classes are illustrated. In the next section, we will use two-dimensional space to show the correlation pattern between two PARCOR parameters among different phoneme classes. For each phoneme class, we will choose more than two samples.



(a) Frame 1, 2 and 3

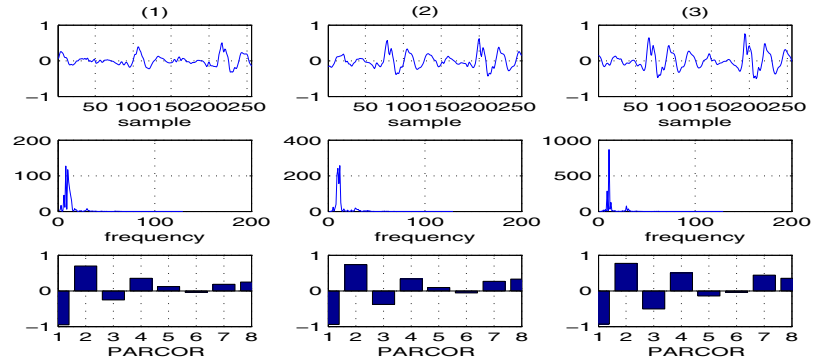


(b) Frame 4, 5 and 6

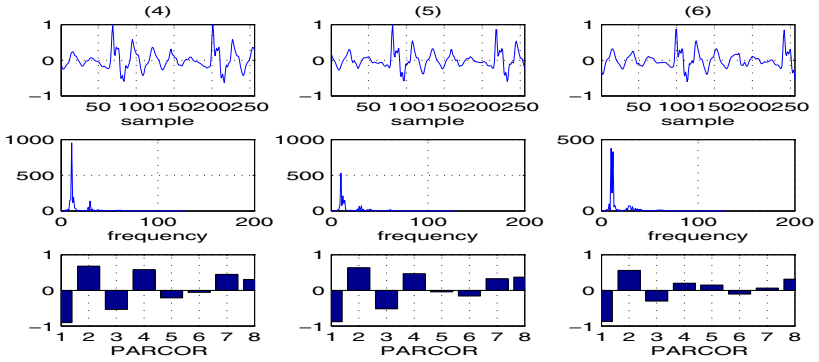


(c) Frame 7, 8 and 9

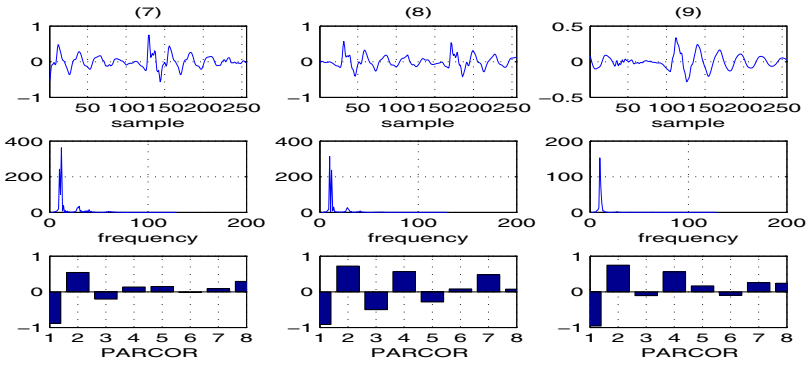
Figure 4.1 Waveforms, spectra and PARCOR distributions of the vowel sound [ae]. Dialect:5 Speaker:female



(a) Frame 1, 2 and 3

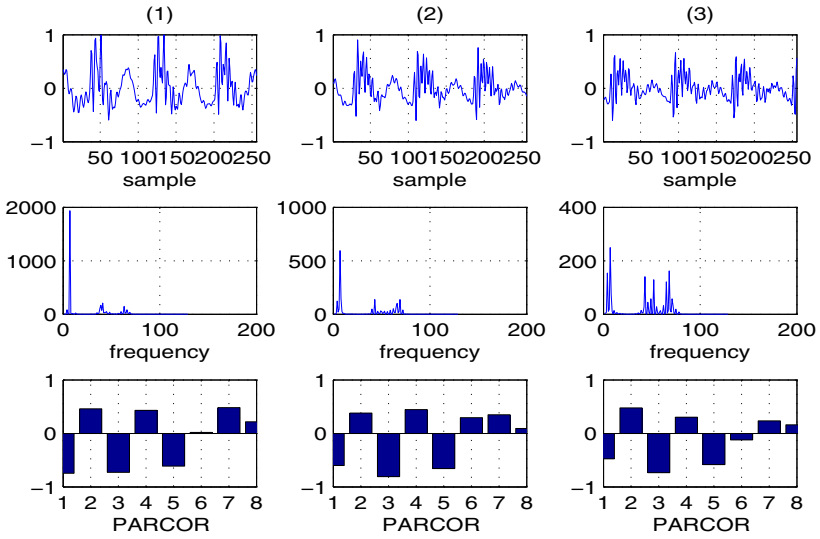


(b) Frame 4, 5 and 6

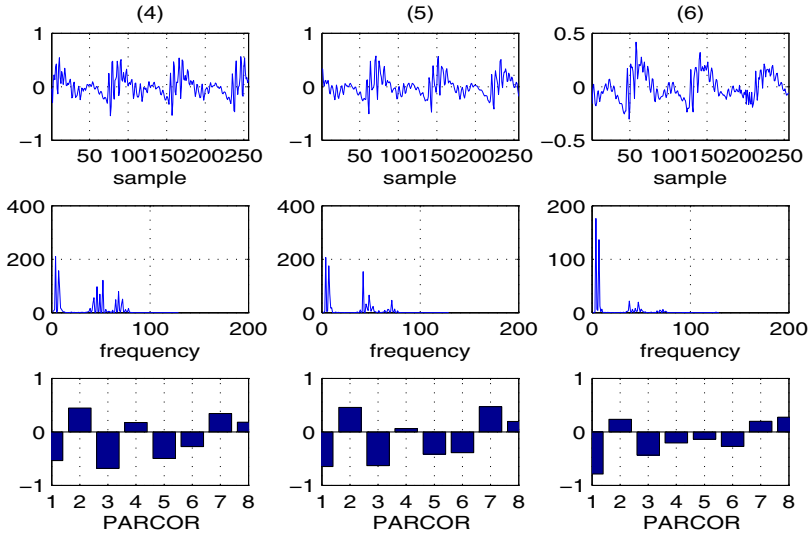


(c) Frame 7, 8 and 9

Figure 4.2 Waveforms, spectra and PARCOR distributions of the vowel sound [ae]. Dialect:4 Speaker:male

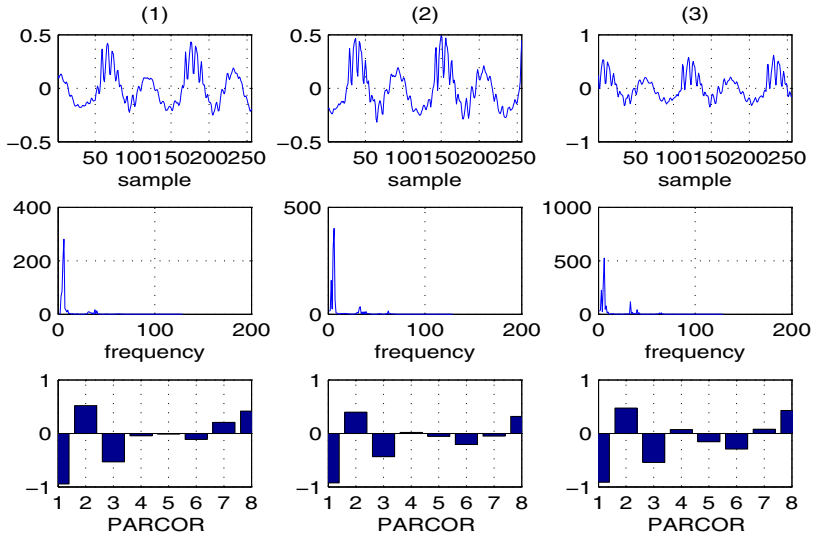


(a) Frame 1, 2 and 3

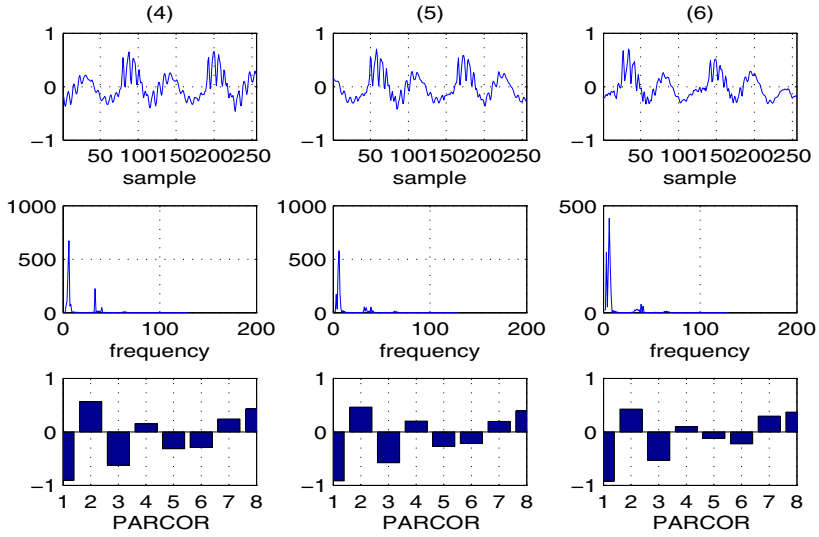


(b) Frame 4, 5 and 6

Figure 4.3 Waveforms, spectra and PARCOR distributions of the vowel sound [iy]. Dialect:3 Speaker:female

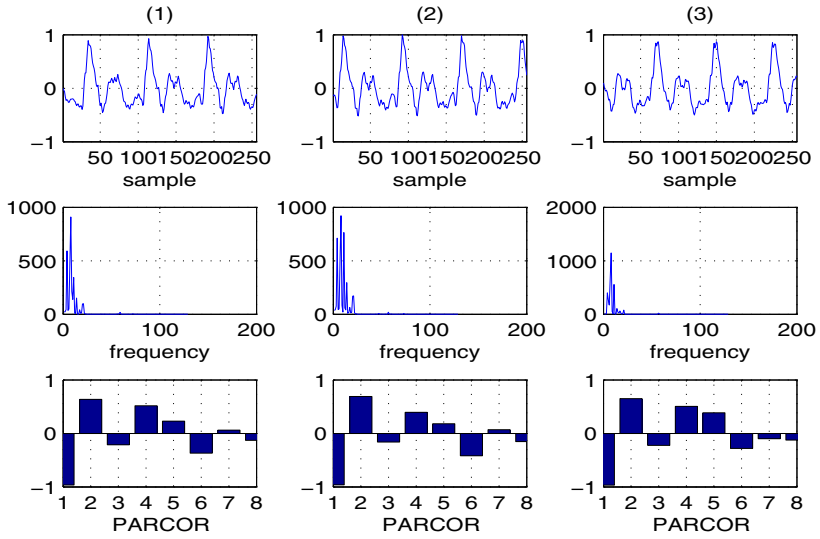


(a) Frame 1, 2 and 3

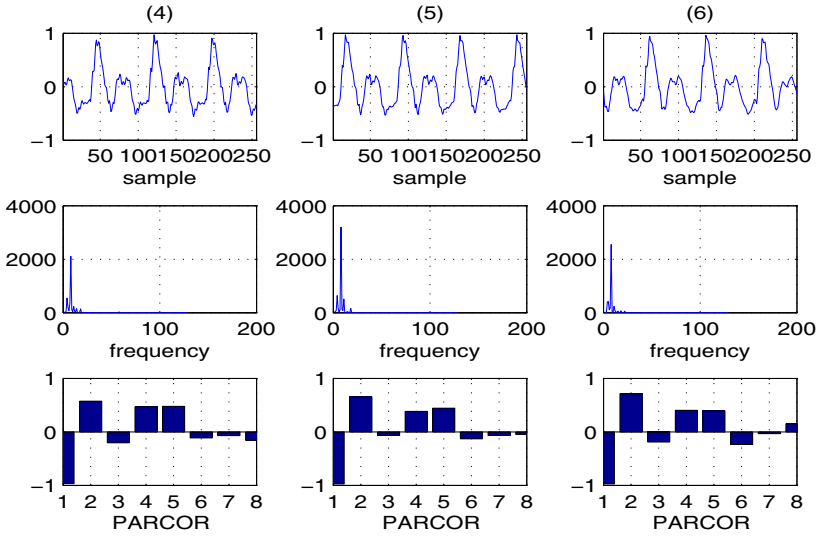


(b) Frame 4, 5 and 6

Figure 4.4 Waveforms, spectra and PARCOR distributions of the vowel sound [iy]. Dialect:4 Speaker:male

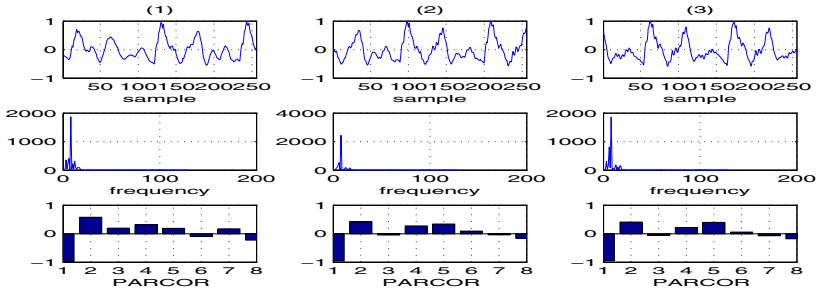


(a) Frame 1, 2 and 3

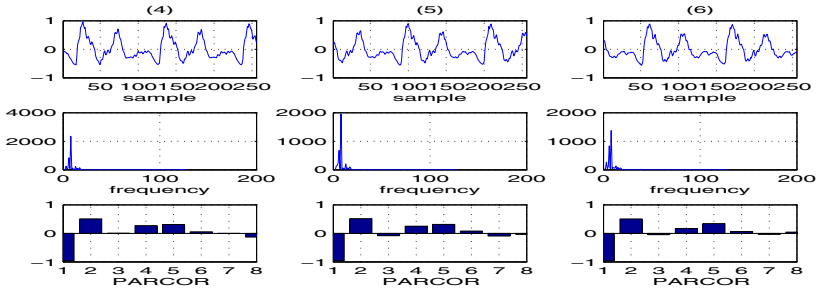


(b) Frame 4, 5 and 6

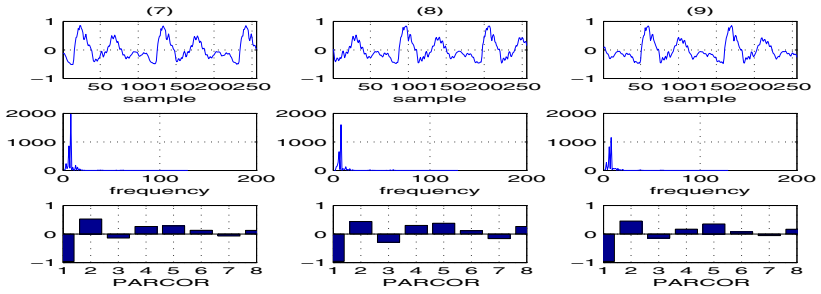
Figure 4.5 Waveforms, spectra and PARCOR distributions of the vowel sound [uw]. Dialect:2 Speaker:female



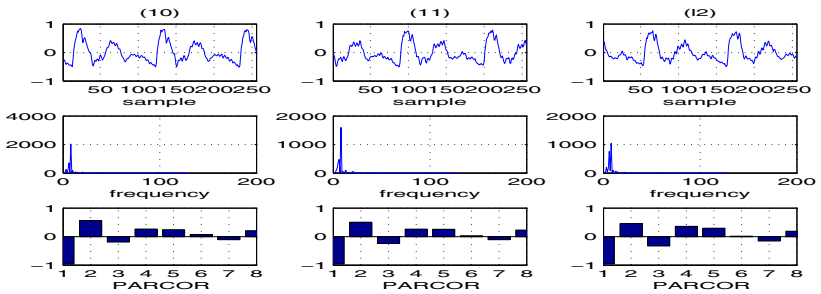
(a) Frame 1, 2 and 3



(b) Frame 4, 5 and 6

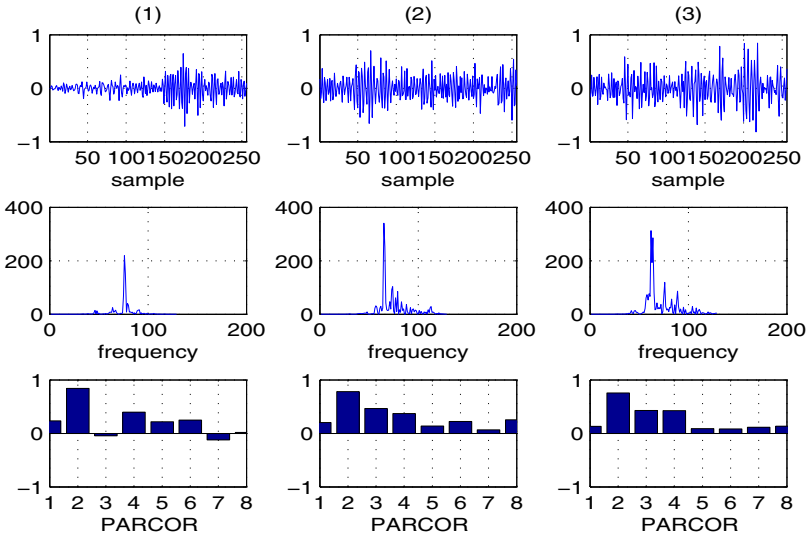


(c) Frame 7, 8 and 9

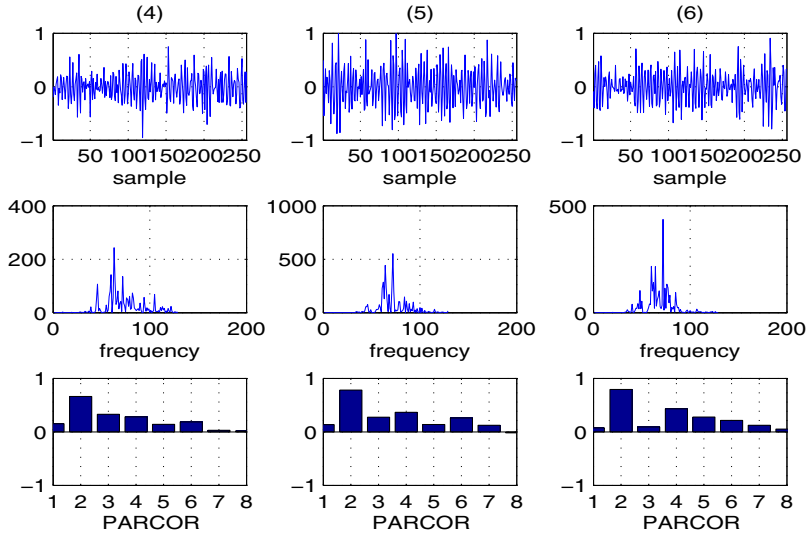


(d) Frame 10, 11 and 12

Figure 4.6 Waveforms, spectra and PARCOR distributions of the vowel sound [uw]. Dialect:6 Speaker:male

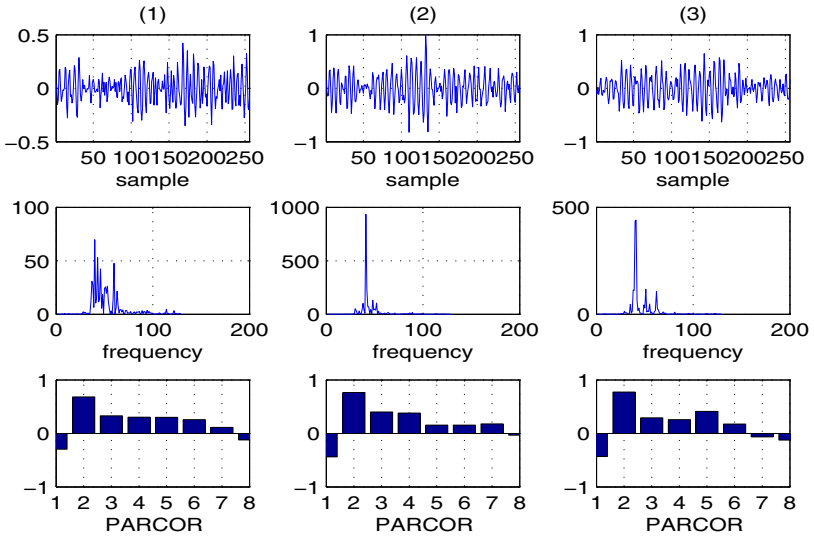


(a) Frame 1, 2 and 3

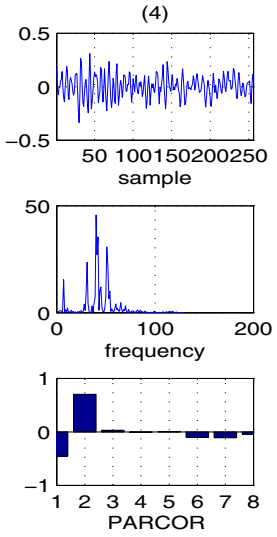


(b) Frame 4, 5 and 6

Figure 4.7 Waveforms, spectra and PARCOR distributions of the consonant sound [sh]. Dialect:4 Speaker:female

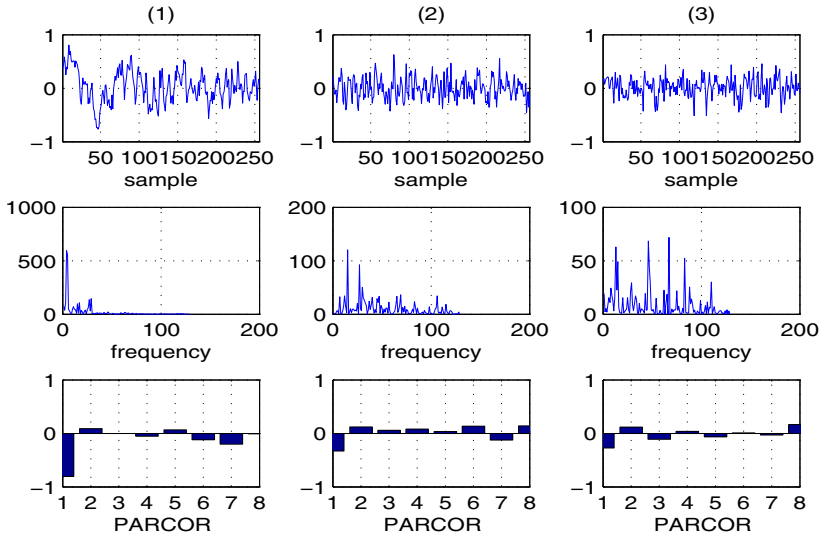


(a) Frame 1, 2 and 3

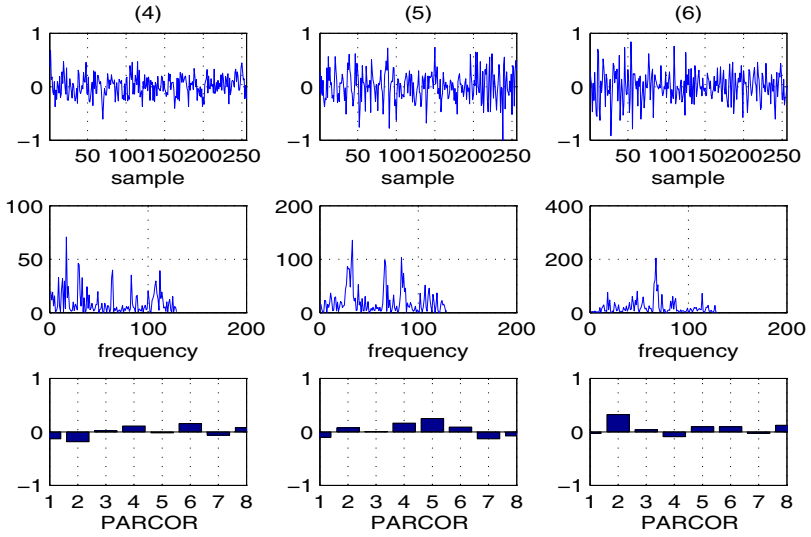


(b) Frame 4

Figure 4.8 Waveforms, spectra and PARCOR distributions of the consonant sound [sh]. Dialect:2 Speaker:male

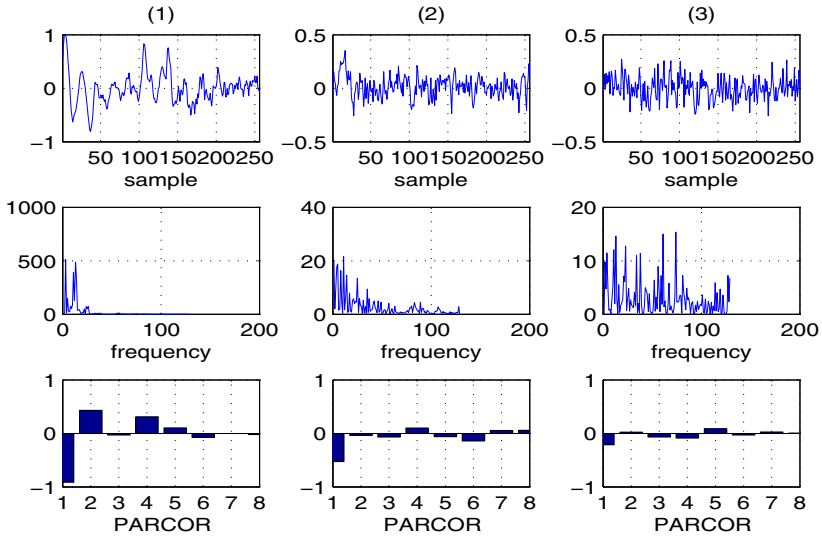


(a) Frame 1, 2 and 3

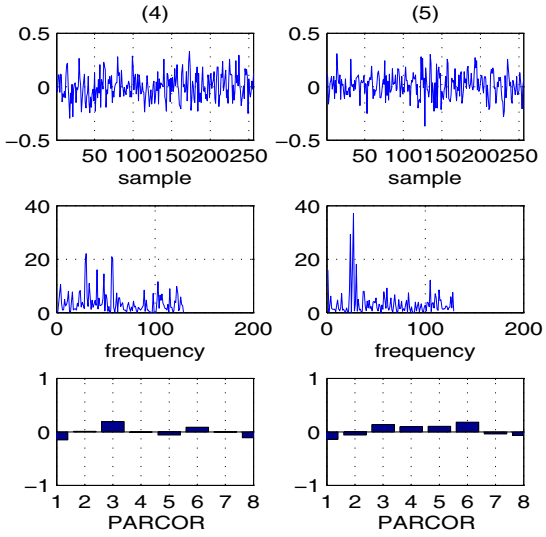


(b) Frame 4, 5 and 6

Figure 4.9 Waveforms, spectra and PARCOR distributions of the consonant sound [f]. Dialect:4 Speaker:female



(a) Frame 1, 2 and 3



(b) Frame 4 and 5

Figure 4.10 Waveforms, spectra and PARCOR distributions of the consonant sound [f]. Dialect:7 Speaker:male

4.2 PARCOR Distributions Among Different Phoneme Classes

In Figure 4.1 to Figure 4.10, we noticed that consecutive 16 ms frames in the same utterance are similar in waveforms, spectra and PARCOR distributions each other. For example, in Figure 4.1, there are total 9 consecutive frames, which are marked (1) to (9) on the top of sub-figures, all nine frames are similar to each other. But we have to mention that, sometimes the first frame and last frame are more different from other frames, as you can seen in Figure 4.3, except the last frame, all other five frame are more similar, in Figure 4.10, the first frame are more different from others. This is because all of the single phoneme sounds are extracted from continuous speech and the first fame and the last frame may be in transition from the previous or to the next frame. In order to show the distributions of PARCOR parameters among different phoneme classes, we select the typical frames from each phoneme class to calculate the corresponding PARCOR parameters, then illustrate the correlation distributions of the eighth-order PARCOR parameters in a two-dimensional space. The distributions of eight PARCOR parameters are divided into two groups of PARCOR parameters. One is the vowel group, the other one is consonant group.

4.2.1 Vowels

First, in TIMIT database we choose three vowels [iy], [ae] and [uw]; then each vowel is segmented into consecutive 256 samples frames. The typical frames are selected from each vowel to calculate the PARCOR parameters. It includes 50 [iy], 50 [ae] and 54 [uw] frames which are spoken by male and female people from eight dialect regions.

PARCOR distributions of typical frames from the vowel [iy], [ae] and [uw] are illustrated in the Figure 4.11. In Figure 4.11, the phoneme [iy], [ae] and [uw] are marked by star, circle and plus respectively. Calculated PARCOR parameters form a cluster for each of [iy], [ae] and [uw] which overlap to a degree, but are separable by properly choosing partitioning lines. Since the database contains same

phoneme spoken by different people from different dialect regions, there are assorted variations, such as time and tone. From Figure 4.11, the potential capability to characterize the phoneme [iy], [ae] and [f] by the PARCOR parameters is indicated.

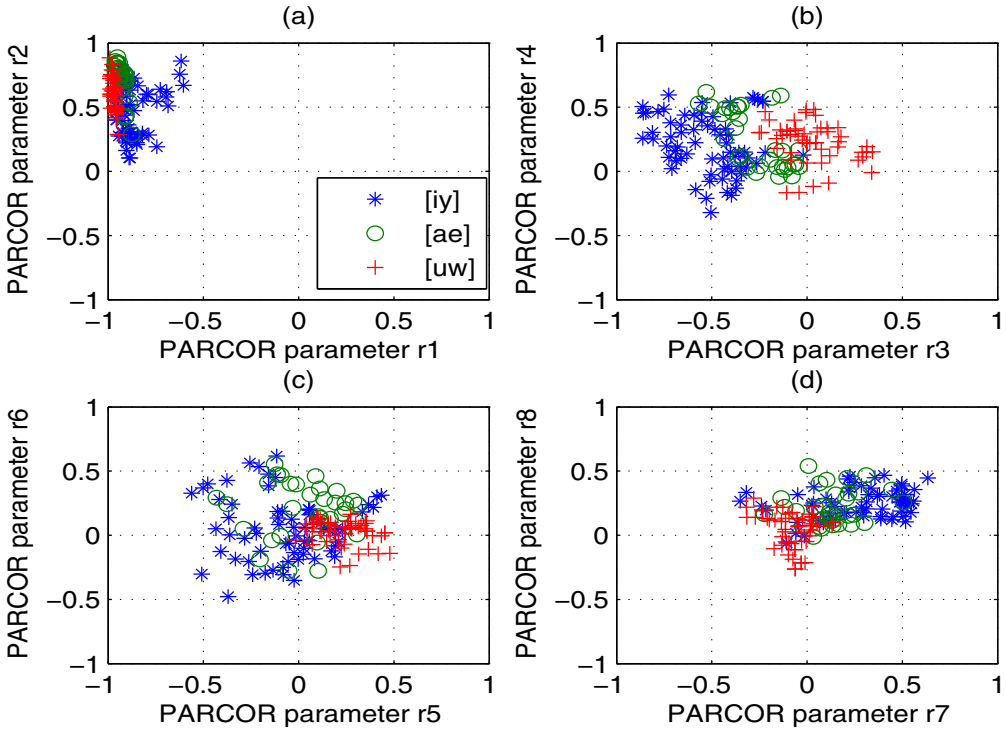


Figure 4.11 Distributions of PARCOR parameters of the vowel [ae], [iy] and [uw]

4.2.2 Consonants

For consonants, we choose the fricative sounds [sh] and [f] in TIMIT database. Similar to the vowel sounds, each consonant is segmented into consecutive 256 samples frames. The typical frames are selected from each consonant to calculate the PARCOR parameters. It includes 101 [sh] and 63 [f] frames which are spoken by male and female people from eight dialect regions.

PARCOR distributions of typical frames from the consonant [sh] and [f] are illustrated in Figure 4.12. In Figure 4.12, the phoneme [sh] and [f] are marked by star and circle respectively. In Figure 4.12 (a), we can see the separation of the cluster of [sh] and [f], the distribution of cluster of [sh] is localized in the upper

region and well separated with [f]. Although there is some degree of overlapping between the cluster [sh] and [f] in Figure 4.12 (b), the separation of the cluster [sh] and [f] appears possible. In Figure 4.12 (c), there is overlap between [sh] and [f], also there is higher degree of overlapping between [sh] and [f] in Figure 4.12 (d), but the separation of consonant [sh] and [f] in Figure 4.12 (a) and (b) is clear, it means the consonant [sh] form a cluster, which is well separated from the cluster of the fricative consonant [f]. From Figure 4.12, the PARCOR parameters have the potential capability to characterize the phoneme [sh] and [f].

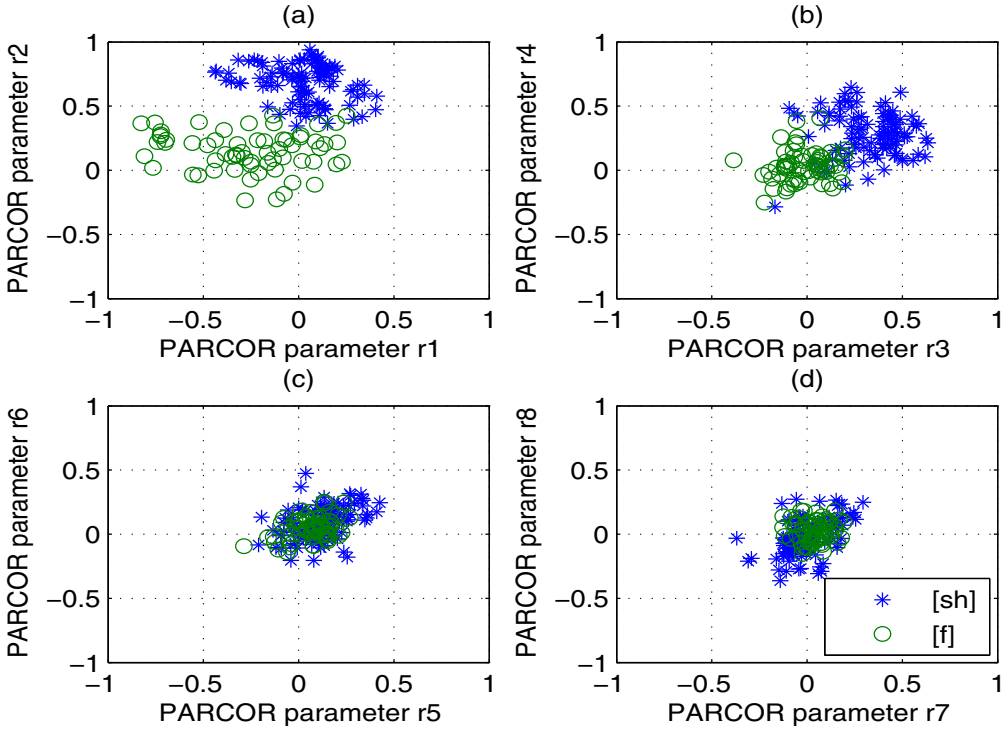


Figure 4.12 Distributions of PARCOR parameters of the consonant [sh] and [f]

Chapter 5

CLASSIFICATION OF PHONEMES

In Chapter 4, we discussed the PARCOR parameters distributions at the phoneme level. In this chapter, we explore a method to classify phonemes in one-syllable words by means of PARCOR parameters in a continuous speech stream.

The phonemes [ae], [iy], [uw], [sh] and [f] in one-syllable words, such as "cat", "greasy", "moon", "she" and "leaf", were chosen to classify. The PARCOR parameters of each phoneme were fed into a classifier, the classifier is a supervised classifier that requires training. The training uses TIMIT speech database, which contains the recordings of 630 speakers of 8 major dialects of American English. The training data were grouped into the vowel group including phoneme [ae], [iy] and [uw] and the consonant group including [sh] and [f]. In the vowel group, there were fifty training data for [ae], fifty training data for [iy] and fifty-four training data for [uw]. The data were selected from eight dialect regions and spoken by different male and female speakers. Similarly, in the consonant group, there were one hundred and one training data for [sh] and sixty-three training data for [f]. They were spoken by different male and female speakers from eight dialect regions.

For the vowel group, including [iy], [ae] and [uw], the eighth-order of PARCOR parameters of each training data were calculated. By observing PARCOR parameters distributions of the vowel training data in a two-dimensional space, which are shown in Figure 5.1, we notice that the cluster of the third and the fourth PARCOR parameters of each vowel is well separated, which is shown in Figure 5.1 (b). When we paid attention to the mean distributions of PARCOR parameters of the vowel training data in a two-dimensional space, which are illustrated in Fig-

ure 5.2, we also notice that the distances between mean of [ae], mean of [iy] and mean of [uw] in Figure 5.2 (b) are bigger than sub-figure (a), (c) and (d). We chose the third and the fourth PARCOR parameters to further process and derive the decision rule. For the consonant group, when we observed the eighth-order of PARCOR parameters distributions of the consonant training data in a two-dimensional space in Figure 5.3, we found that the third and the fourth PARCOR parameters of each consonant clustered together better than others. We also selected the third and fourth PARCOR parameters to further train and process. Assuming the two PARCOR parameters (the third and the fourth of the eighth-order PARCOR parameters) of each phoneme sound are Gaussian distributions, we can construct a Gaussian distribution template for each phoneme sound. The Gaussian probability distribution templates of vowel [ae], [iy] and [uw] are constructed by using training data in the vowel group. For the consonant group, we construct two Gaussian probability distribution templates, one is for [sh] and the other one is for [f]. The training data in the consonant group are used to construct the probability distribution.

In order to classify the unknown phonemes in one-syllable word into the [ae], [iy], [uw], [sh] or [f] class, we designed two classifiers, one is a vowel classifier and the other one is a consonant classifier. For the vowel classifier, the unknown phoneme can be classified into one of [ae], [iy] and [uw] classes. For the consonant classifier, the unknown phoneme can be classified into either [sh] or [f]. For both classifiers, the maximum likelihood decision rule is adopted to classify the unknown phoneme. That is, when the third and fourth PARCOR parameters are input into the classifier, the classifier will calculate and compare the probability of each phoneme and then decide the unknown phoneme belongs to the phoneme class which has the maximum probability. For instance, when the third and fourth PARCOR parameters of an unknown phoneme are fed into the vowel classifier, the vowel classifier will calculate the probability of vowel [ae], [iy] and [uw] respectively and compare the three values of the probabilities, if the probability of vowel [iy] has the maximum value, then the unknown phoneme will be classified into the vowel [iy] class. Applying this procedure to different phonemes, we can classify the unknown phonemes into [ae],

[iy] or [uw] class.

Since the input of the both classifiers are the third and the fourth PARCOR parameters of unknown vowel or consonant phonemes, we need to explore a method to preprocess unknown phonemes and calculate their corresponding PARCOR parameters. This method also is accountable for detecting the vowel and consonant phonemes in one-syllable words in order to feed the PARCOR parameters into either the vowel classifier or the consonant classifier. The preprocessing method is illustrated in Figure 5.9. The method is broadly divided into three steps, the first step is to segment speech signals by frame energy and zero-crossing rate, then group frames into consonant, vowel or silence, the last step is to calculate the PARCOR parameters for the vowel and consonant group and the third and the fourth PARCOR parameters are selected to feed into the classifier. The calculated third and fourth PARCOR parameters of unknown phonemes from the vowel group were fed into the vowel classifier and The calculated the third and the fourth PARCOR parameters of unknown phonemes from the consonant group were fed into the consonant classifier.

5.1 Training and Derivation of the Decision Rule

The training data are divided into the vowel group and the consonant group. In the vowel group, the phoneme [ae], [iy] and [uw] are selected from eight dialect regions and spoken by different male and female speakers. There are fifty training data for [ae], fifty training data for [iy] and fifty-four training data for [uw]. All of vowel phonemes are segmented into consecutive 16 ms frames and typical frames are selected from each vowel to calculate the eighth-order of PARCOR parameters. PARCOR parameters distributions of the vowel training data in a two-dimensional space are shown in Figure 5.1. In Figure 5.2, it shows the corresponding mean distributions of PARCOR parameters of the vowel training data in a two-dimensional space. In Figure 5.1, we can see the the phoneme [iy], [ae] and [uw], which are marked by star, circle and plus respectively, form cluster, particularly in sub-figure (b) of Figure 5.1, the cluster [iy], [ae] and [uw] are separated better than other

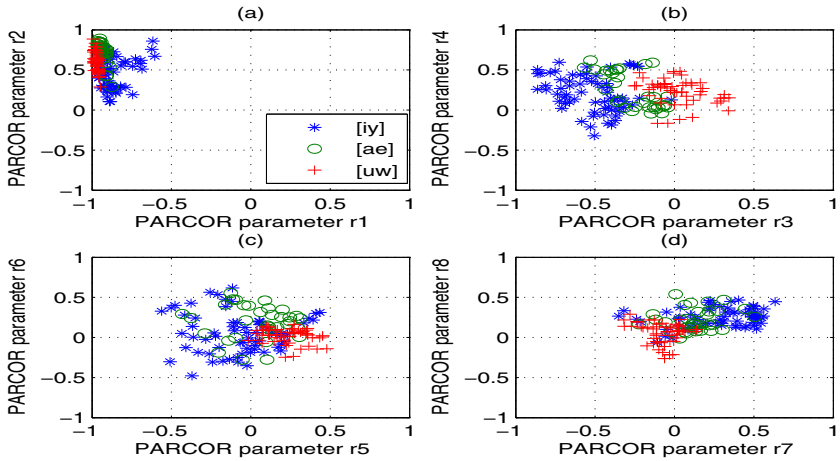


Figure 5.1 PARCOR parameters distributions of the vowel training data in a two-dimensional space

sub-figure (a), (c) and (d). When we turn our attention to Figure 5.2, we observe that the mean distributions of PARCOR parameters of the vowel [iy], [ae] and [uw] are separated better in sub-figure (b) than other sub-figures. We selected the third and the fourth PARCOR parameters as feature vector to further process and derive the decision rule.

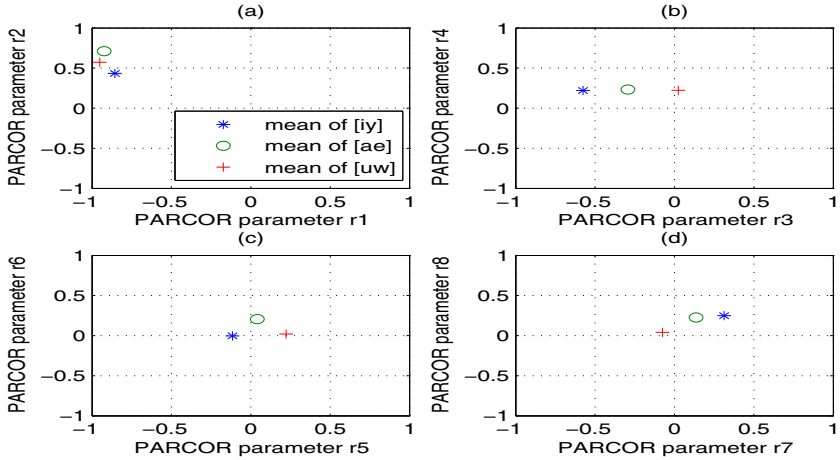


Figure 5.2 Mean distributions of PARCOR parameters of the vowel training data in a two-dimensional space

In the consonant group, it includes the phoneme [sh] and [f] and the training phonemes are selected from eight dialect regions and spoken by different male and

female speakers. There are one hundred and one for [sh] and sixty-three for [f], which are spoken by different male and female speakers from eight dialect regions. All consonant phonemes are segmented into consecutive 16 ms frames and typical frames are selected from each consonant to calculate the eighth-order of PARCOR parameters. PARCOR parameters distributions of the consonant training data in a two-dimensional space are shown in Figure 5.3. In Figure 5.4, the corresponding mean distributions of PARCOR parameters of the consonant training data in a two-dimensional space are illustrated. In Figure 5.3, we can see the the phoneme

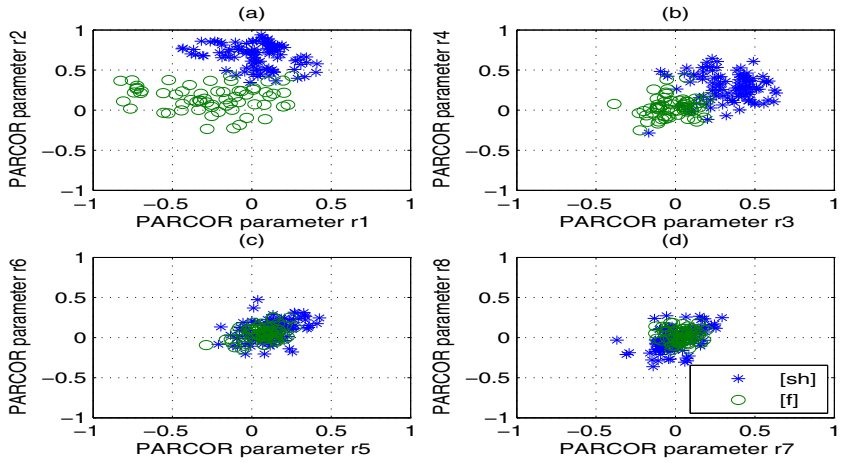


Figure 5.3 PARCOR parameters distributions of the consonant training data in a two-dimensional space

[sh] and [f], which are marked by star and circle respectively, form separable cluster, particularly in sub-figure (a) and (b), the cluster of [sh] and [f] are separated better than sub-figure (c) and (d). Comparing sub-figure (a) with sub-figure (b) in Figure 5.3, the cluster of [f] more condense in sub-figure (b) than sub-figure (a). When we observed Figure 5.4, we noticed that the mean of [sh] and the mean of [f] are separated very well in sub-figure (b), the third and the fourth PARCOR parameters are selected as feature vector to further process in the consonant group.

Assuming distributions of the third and fourth PARCOR parameters in both vowel and consonant group are Gaussian distributions, the Gaussian distribution probability of PARCOR parameters can be estimated by using the sample data in training set for each phoneme class. The following equations were used to construct

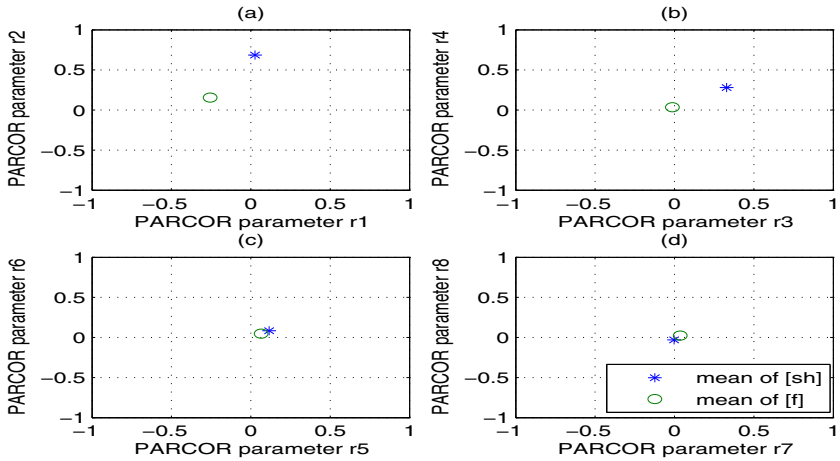


Figure 5.4 Mean distributions of PARCOR parameters of the consonant training data in a two-dimensional space

the Gaussian density function and estimate the statistics parameters, such as mean and variance. In the Equation 5.1, the density function of a multivariate Gaussian is illustrated. [21]

$$f_x(x) = \frac{1}{(2\pi)^{n/2}|C_x|^{1/2}} e^{-\frac{1}{2}(x-m_x)^T C_x^{-1}(x-m_x)} \tag{5.1}$$

where n is the dimension of x. This density function is completely characterized by the mean vector m_x and the covariance matrix C_x, which are given in Equation 5.2 and Equation 5.3.

$$m_x = E\{x\} = \int_x x f_x(x) dx \tag{5.2}$$

$$\begin{aligned} C_x &= E\{(x-m_x)(x-m_x)^T\} \\ &= \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \vdots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} \end{aligned} \tag{5.3}$$

How to estimate the mean and covariance from samples in training data set is given

by Equation 5.4 and Equation 5.5. [27]

$$m_x \approx \frac{1}{M}\sum_{j=1}^{M} y_j \tag{5.4}$$

$$\sigma_{ij} \approx \frac{1}{M}\sum_{k=1}^{M}(y_{ki} - m_i)(y_{kj} - m_j) \tag{5.5}$$

where the M is the number of samples.

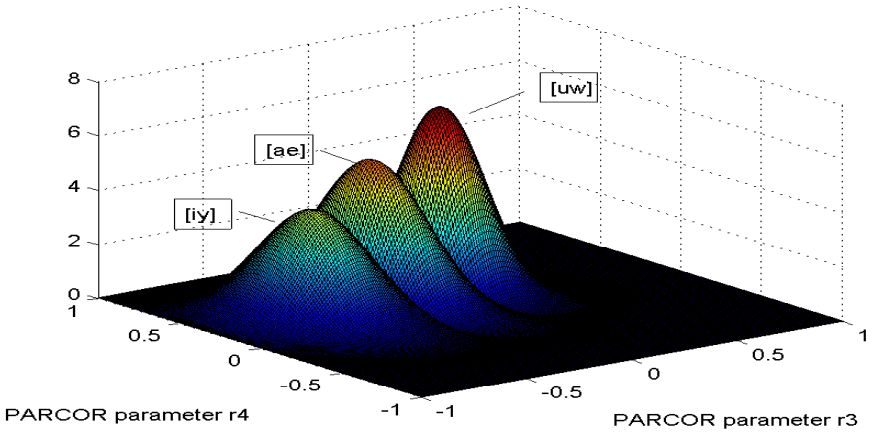


Figure 5.5 Estimated Gaussian density functions of PARCOR parameters of the vowel [iy], [ae] and [uw]

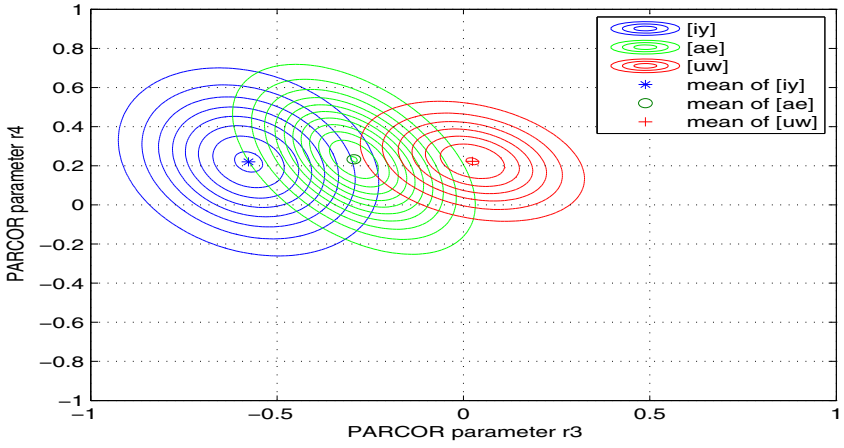


Figure 5.6 Contour lines of estimated Gaussian density functions of the vowel [iy], [ae] and [uw]

For the vowel group, we construct the Gaussian density function of the third and fourth PARCOR parameters for each vowel [iy], [ae] and [uw] by using the samples in training data set and Equation 5.1 to Equation 5.5. The estimated Gaussian density functions of PARCOR parameters of the vowel [iy], [ae] and [uw] are illustrated in Figure 5.5. The corresponding contour lines of the estimated Gaussian density functions are give in Figure 5.6. Similarly, for the consonant group, we construct the Gaussian density function of the third and fourth PARCOR parameters for each consonant [sh] and [f] by using the samples in training data set and Equation 5.1 to Equation 5.5. In Figure 5.7, it shows the estimated Gaussian density functions of PARCOR parameters of the consonant [sh] and [f]. The corresponding contour lines of the estimated Gaussian density functions are give in Figure 5.8.

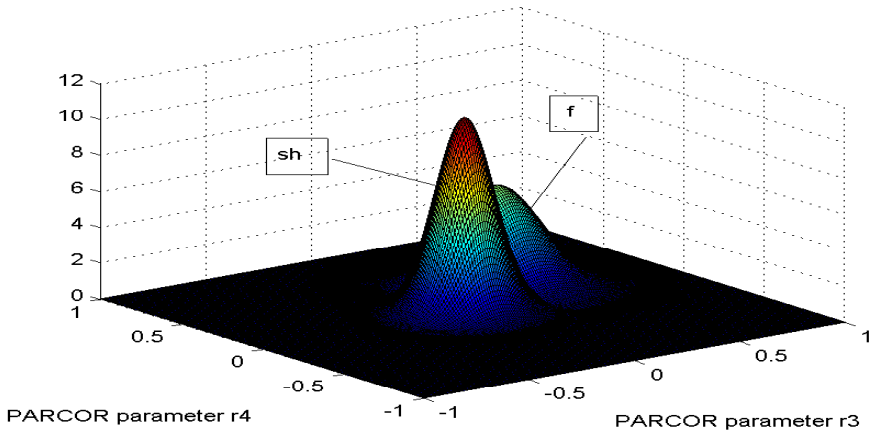


Figure 5.7 Estimated Gaussian density functions of PARCOR parameters of the consonant [sh] and [f]

By observing Figure 5.5 to Figure 5.8, we found that each phoneme class in both vowel and consonant group can be partitioned by choosing an appropriate partition line. The maximum likelihood decision rule is selected to determine partition line. The maximum likelihood decision rule (MLDR) is given by Equation 5.6. [27]

$$C(x) = j \quad if \quad p_j p_j(x) \geq p_k p_k(x) \quad for \quad all \quad k \neq j \tag{5.6}$$

where $C(x)$ is defined as the decision function, the value of which is the best choice

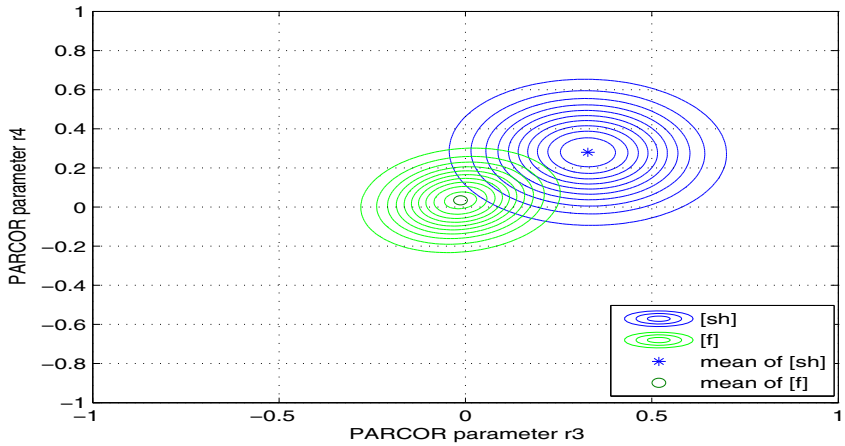


Figure 5.8 Contour lines of estimated Gaussian density functions of the consonant [sh] and [f]

of class to which to assign x. Where p_j is the apriori probability and $p_j(x)$ is the class conditional probability. The j and k denote different classes. In principle, p_j is determined by some knowledge that is independent of the observation of samples. Usually the apriori probability is supposed to be provided by designer's experience and knowledge. The other key part of maximum likelihood decision rule is to estimate the class conditional probability. For each vowel and consonant class in our training data set, we assume the apriori probabilities are equal. The class conditional probabilities are constructed from the samples in training data set by means of Equation 5.1 to Equation 5.5. The construed class conditional probabilities for [ae], [iy], [uw], [sh] and [f] are shown in Figure 5.5 and Figure 5.7.

5.2 Classification

5.2.1 Data Conditioning for Classification

In this preprocessing for classification, we extract the one-syllable word from a continuous stream of speech in TIMIT database. [13] Then, we segmented the word into consecutive frames and extract typical vowel and consonant frames to calculate the corresponding PARCOR parameters. The preprocessing method can be broadly divided into the following steps shown in Figure 5.9.

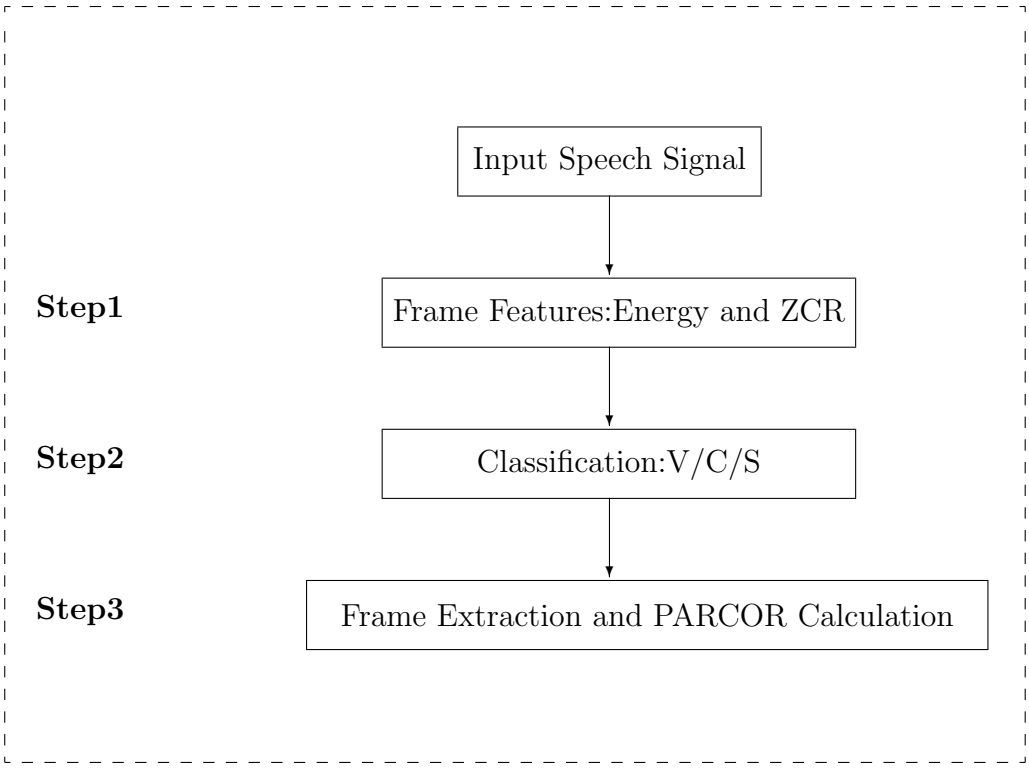


Figure 5.9 Illustration of preprocessing method

Segmentation of Speech Signals by Frame Energy and Zero-Crossing Rate

In the first step, basic features are extracted. The input speech signals are segmented into 16 ms long, non-overlapping frames. Two frame features, frame energy and zero-crossing rate (ZCR), are calculated. The energy and ZCR are used to segment speech into smaller units corresponding to phonemes and to find the boundary between voiced and unvoiced portions of speech.

The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. In general, we can define the short-time energy as [1]

$$E_n = \sum_{m=-\infty}^{m=+\infty} [x(m)w(n-m)]^2 \tag{5.7}$$

where $w(n-m)$is the window function. In our experiments, we choose the rectangular window as

$$w(n) = 1;\ 0 \leq n \leq N-1$$

$$w(n) = 0;\ otherwise \tag{5.8}$$

The choice of the window determines the nature of the short-time energy representation. The value of N can not be too large or small. If N is too large En will change very slowly and thus will not adequately reflect the changing properties of the speech signals in energy. If N is too small, En will fluctuate very rapidly. We chose N equal 256. Since the speech signal in TIMIT database has the 16K sample rate, the 256 consecutive samples frame is 16 ms long in duration.

In Figure 5.10, we show the energy of word "cat", which is extracted from in the continuous speech "Get a calico cat to keep" and spoken by a female person who is from dialect region five. The word "cat" is segmented into consecutive 16 ms frames and the energy of each frame is calculated by Equation 5.7 and Equation 5.8. For high quality speech (high signal-to-noise ratio), the En can be used to distinguish speech from silence. From Figure 5.10, it is clear that frames segmented from the

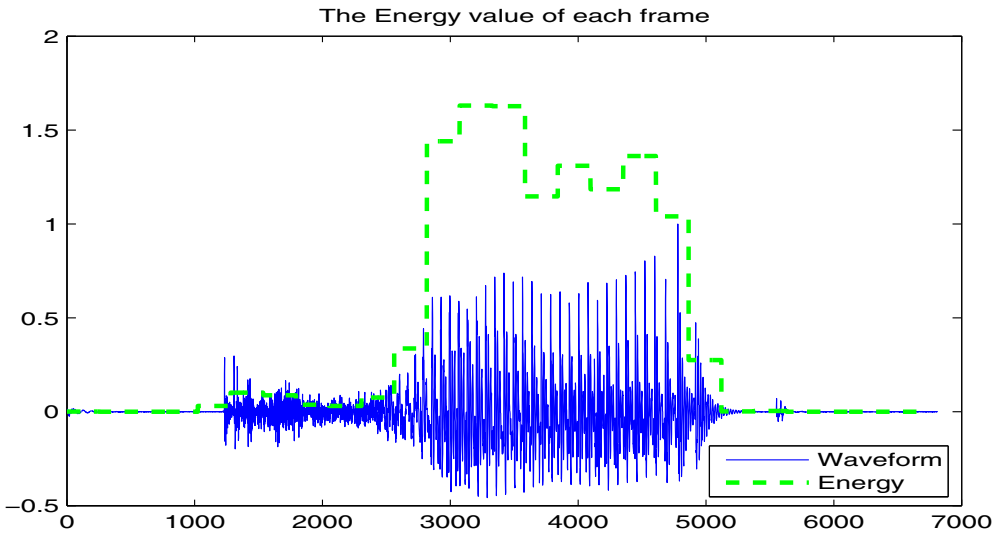


Figure 5.10 Energy value of each frame in the word "cat"

vowel [ae] have the high energy. The frames at the beginning and the end of the word have the low energy, which are consonant [k] and [t].

The other selected frame feature is ZCR. ZCR stands for Zero Crossing Rate, it occurs if successive samples have different algebraic signs. The value of ZCR is

defined by the following equation [1]

$$ZCR = \sum_{m=-\infty}^{m=+\infty} |sgn[x(m+1)] - sgn[x(m)]|\, w(n-m) \tag{5.9}$$

where sgn function is given by

$$\begin{aligned} sgn[x(m)] &= 1; \quad x(m) > 0 \\ sgn[x(m)] &= 0; \quad x(m) = 0 \\ sgn[x(m)] &= -1; \quad x(m) < 0 \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} w(n) &= \tfrac{1}{2}; \quad 0 \le n \le N-1 \\ w(n) &= 0; \quad otherwise \end{aligned} \tag{5.11}$$

Equation 5.9 to Equation 5.11 means that the ZCR is to check samples in pairs and find where the zero-crossings occur.

We show the ZCR of each frame in a word "cat" in Figure 5.11. In Figure 5.11,

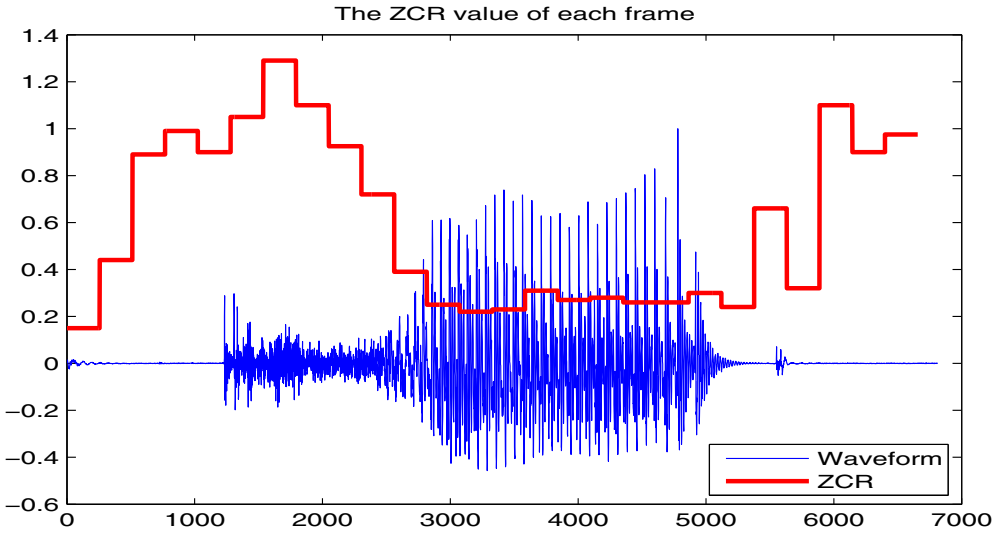


Figure 5.11 ZCR value of each frame in the word "cat"

we can see ZCR values of the vowel frames are relatively low. But the consonant frames with high ZCR value are noticeable in the beginning and end of the word

"cat". Generally, we draw the conclusion that if the zero-crossing rate is high, it indicates the speech is unvoiced, and if the zero-crossing rate is low, it indicates that the speech is voiced speech. So ZCR is quite useful in making the distinction between voiced and unvoiced speech.

From speech production, we know that voiced speech produced by the quasi-periodic air to excite the relatively fixed vocal tract, so the energy of voiced speech is concentrated below about 3 KHz , whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero-crossing rate, and low frequencies imply low zero-crossing rate, there is a correlation between zero-crossing rate and energy distribution with frequency, we illustrate it in the Figure 5.12.

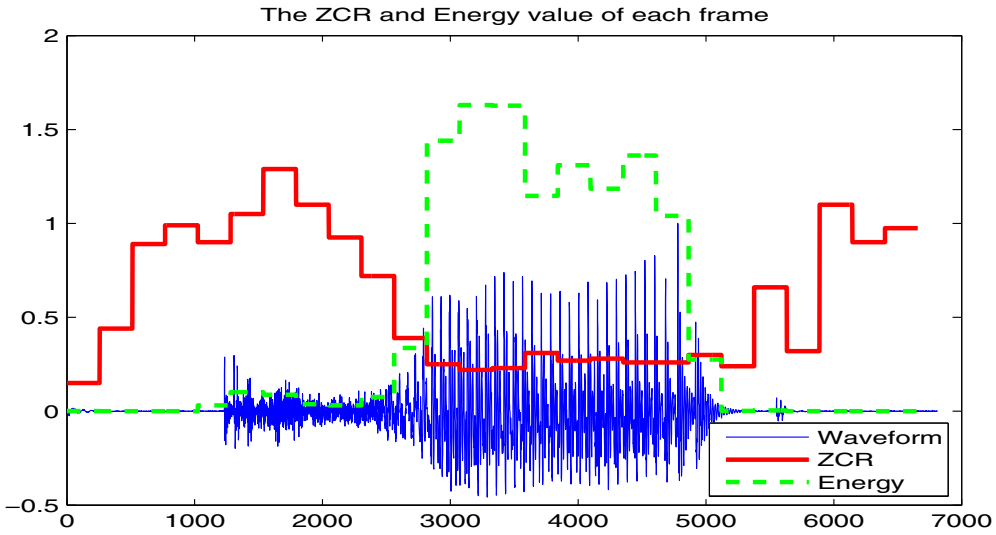


Figure 5.12 Correlation of energy and ZCR value of each frame in the word "cat"

Grouping Frames into Consonant, Vowel or Silence

In the second step, all consecutive frames, which segmented by the rectangular window which is shown in Equation 5.8, are grouped into C/V/S (Consonant/Vowel/Silence) at the phoneme level. By using the following rules, frames are classified to one of C/V/S groups. We already know ZCR provides a basis for distinguishing voiced speech frames from unvoiced speech frames, so a threshold of

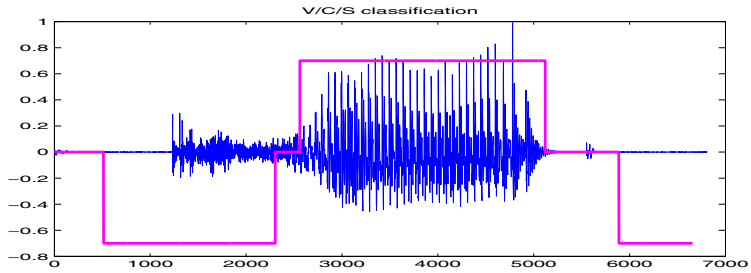
ZCR should be set up. In this experiment, the threshold of ZCR is set by the mean of ZCR plus the standard deviation of ZCR divided by two. If one frame's ZCR value is greater than the threshold of ZCR, it will be determined to be a consonant frame. Then we set a threshold of energy for distinguishing the silence frame from vowel frames. The threshold of energy is given by the mean of energy plus the standard deviation of energy divided by two. If the frame's energy is lower than or equal to the threshold of energy, it will be decided into the silence frame group. The rest of the frames, whose ZCR values are less than or equal to the threshold of ZCR and energy is greater than or equal to the threshold of energy, all belong to the vowel frames. The rules can be summarized in the following:

- Compare the frame's ZCR value with the threshold of ZCR, if ZCR is greater the threshold of ZCR, the frame will be grouped into the consonant frame group, otherwise the frame's energy will be calculated.
- Compare energy of the frame, whose ZCR is less than and equal to ZCR threshold with the threshold of energy, if energy is less the energy threshold, then the frame is determined as a silent frame.
- The rest of frames belong to vowel frames.

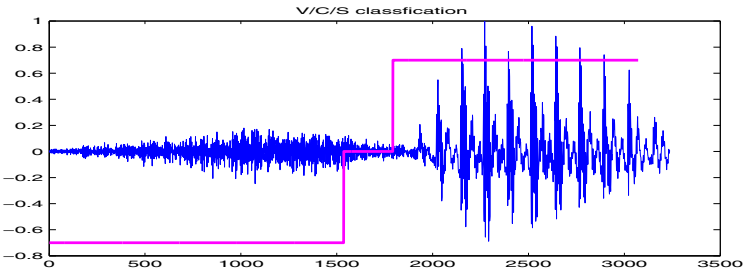
We applied the rules above to the word "cat" and "she", we can classify the vowel, consonant and silence frames, the results are shown in Figure 5.13. The positive values means vowel frames, negative values are consonant and zero are silence frames.

Calculating PARCOR Parameters

In the third phase, the typical vowel and consonant frames are extracted in order to calculate the PARCOR parameters. We search for a contiguous stretch of vowel frames, then select the center frame of the vowel group as a typical vowel frame. Once the typical vowel frame is located, the consonant frames are determined by backward and forward search from the position of the typical vowel frame. The first on-set of a contiguous vowel frame group and the last frame of the group are



(a) Classification of each frame in the word "cat"



(b) Classification of each frame in the word "she"

Figure 5.13 Classification of each frame in the word "she" and "cat"

tested in their similarity to the typical vowel. If similar, the search moves forward and backward to find a frame which is different from the typical vowel frame but still has sufficient energy.

Now, the PARCOR parameters can be calculated for the selected typical vowel and consonant frame. Here the eighth-order PARCOR parameters are calculated. It is clear that if the order of the PARCOR parameters is higher, more details of the speech characteristics are represented. But our objective is to characterize speech signals by using PARCOR parameters at a considerably lower information rate to realize speech recognition. The eighth-order PARCOR parameters are appropriate to reflect the characteristics of speech signals. The Autocorrelation method was implemented to solve the PARCOR parameters. When this method is applied, the large prediction error occurred at the beginning of the frame due to the fact that we are attempting to predict the samples of the signals from the zero valued samples

outside the interval.

5.2.2 Classification Results and Discussion

In TIMIT database, some one-syllable words are selected as the test words. The test words are chosen from eight dialect regions and are spoken by different male and female speakers, the information of the test words, such as speaker name, gender and dialect region, are listed in Appendix B.

Based on the method, which we talked in Section 5.2.1, test words were pre-processed for the classification. After the preprocessing, the PARCOR parameters of typical vowel and consonant frames were obtained. Then the third and fourth PARCOR parameters are fed into either the vowel classifier or consonant classifier. Finally, we classify the phonemes by using the maximum likelihood decision rule, which is introduced in Section 5.1. The vowel and consonant classification results are shown in Table 5.1 to Table 5.7. The summary of the results are listed from Table 5.8 to Table 5.14.

In Table 5.1 and Table 5.2, the classification results of the phoneme [iy] in the one-syllable word are listed. There are totally fifty test words, which are selected from eight dialect regions. We noticed that the classification rate is different in different dialect regions, as is summarized in Table 5.8. Similarly, we summarized the classification results of vowel [ae] and [uw] in different dialect regions in Table 5.9 and Table 5.10. In Table 5.11, it shows the vowel [ae], [iy] and [uw] classification rate respectively and the total classification rate of the vowel group. In Table 5.8, we can see the total eleven test words in dialect region one and eight are all classified correctly. The classification results for vowel [iy] in dialect region two has the lowest correct classification rate, which is 54.55%. In Table 5.9, we can see the highest correct classification rate of 75.00% is in dialect region one and three, there are four test words in each dialect region and each account for 8.16% of all the test words. In dialect region two and six, there are ten test data respectively, which totally account for 40.82% of all the test words, both of the correct classification rates are 60.00%. In Table 5.10, the correct classification rate is same in dialect region one,

three, four, six and seven, all of them are 60.00%. The lowest correct classification rate is 57.14% in dialect region two. In Table 5.11, it shows the summary of vowel [ae], [iy] and [uw] classification rate respectively and the total classification rate in the vowel group. The correct classification rate of the vowel [iy] is 74.00% for all test words from eight dialect regions. Test words for the vowel [iy] are totally 50, which account for 36.23% of all test words in the vowel group. The correct classification rate of the vowel [ae] is 63.27% for all test words from eight dialect regions. Test words for the vowel [ae] are totally 49, which account for 35.51% of all test words in the vowel group. The correct classification rate of the vowel [uw] is 58.97% for all test words from eight dialect regions. Test words for the vowel [uw] are totally 39, which account for 28.26% of all test words in the vowel group. In total, there are 138 test words in the vowel [iy] [ae] [uw] group, and the correct classification rate is 65.22%.

For consonant [sh] and [f], we summarized the classification results of consonant [sh] and [f] in different dialect regions in Table 5.12 and Table 5.13. In Table 5.14, we can see consonant [sh] and [f] classification rate respectively and the total classification rate in the consonant group. In Table 5.12, except dialect region one and two, the test words are all classified correctly in other dialect regions. The correct classification rate in dialect region one and dialect region two are 75.00% and 60.00% and the number of test words account for 20.00% and 12.50% respectively. From Table 5.13, we can see the correct classification rate is very high in all dialect regions, the correct classification rate is 100.00% from dialect region one to seven, only in dialect region eight, the correct classification rate drops to 80.00%, there are five test words in this dialect region and account for 13.51% of all test words. In Table 5.14, it shows the summary of consonant [sh] and [f] classification rate respectively and the total classification rate in the consonant group. The correct classification rate of the consonant [sh] is 90.00% for all test words from eight dialect regions. Test words for the consonant [sh] are totally 40, which account for 51.95% of all consonant test words. The correct classification rate of the consonant [f] is 97.30% for all test words from eight dialect regions. Test words for the consonant

[f] are totally 37, which account for 48.05% of all test words in consonant group. In total, there are 77 test words in the consonant group, and the correct classification rate is 93.51% for the consonant group.

The consonant group has the higher correct classification rate (93.51%) than the vowel group (65.22%), we can explain it from the the following facts, one is that there are only two classes in the consonant group, but there are three classes in the vowel group, from the statistics view, the probability of correct classification of unknown consonant is 50.00% and 33.33% for unknown vowel. The other fact, we can see from Figure 5.6 and Figure 5.8, the overlap degree between vowel [iy], [ae] and [uw] is bigger than [sh] and [f].

But for both group, the vowel and the consonant group give us a good enough classification result, particularly for consonant group. In this thesis, the eighth-order PARCOR parameters are calculated, then the four combination of eight PARCOR parameters are investigated in the training stage, finally the third and the fourth PARCOR parameters are selected as feature vector to classify the unknown phonemes. We know the combination of eight PARCOR parameters is twenty-eight. It further proves that the PARCOR parameters have the potential ability to classify the unknown phonemes.

Table 5.1 Vowel [iy] classification results (Test words from dialect region 1, 2, 3, 4 and 5)

Test Word	[iy]	[ae]	[uw]	Results	True/False
She	0.3129	0.0624	3.4422e-005	[iy]	T
She	1.3950	0.7121	0.0034	[iy]	T
Me	0.9030	0.0370	7.2836e-006	[iy]	T
Seem	1.8390	0.0047	1.6732e-006	[iy]	T
Be	2.0656	0.0126	6.1008e-006	[iy]	T
Seed	1.4525	0.3409	8.3558e-004	[iy]	T
She	2.6460	0.0412	2.4901e-005	[iy]	T
She	0.0228	0.5543	3.1422	[uw]	F
He	1.2441	1.6718	0.0134	[ae]	F
Clean	0.2098	1.7719	0.3786	[ae]	F
Street	3.2695	0.0668	5.8643e-005	[iy]	T
Feel	0.0323	1.3110	1.2176	[ae]	F
Leaf	3.2226	0.0508	4.1710e-005	[iy]	T
Each	3.9527	0.2703	5.0105e-004	[iy]	T
Real	1.7551	2.5096	0.0332	[ae]	F
Cream	2.1624	0.0080	3.5465e-006	[iy]	T
Me	2.5010	0.0116	5.9452e-006	[iy]	T
Feed	3.9742	0.6788	0.0022	[iy]	T
She	3.3691	0.0687	7.2246e-005	[iy]	T
Me	2.4032	0.6576	0.0027	[iy]	T
Tea	0.0291	0.8880	0.1611	[ae]	F
Me	3.6870	1.4478	0.0101	[iy]	T
Steep	3.7542	0.1358	1.9426e-004	[iy]	T
Read	0.0063	0.1524	2.1202	[uw]	F
She	1.7835	2.9313	0.0506	[ae]	F
Me	3.5686	0.0883	9.5784e-005	[iy]	T
Be	3.6553	0.1442	1.8347e-004	[iy]	T
Beach	1.6908	0.0111	3.3649e-006	[iy]	T
Flee	1.9879	0.0045	1.7004e-006	[iy]	T
Seem	2.4966	1.5051	0.0085	[iy]	T
See	1.7310	4.6623	0.2045	[ae]	F
Cheap	0.3118	3.6082e-004	1.9246e-008	[iy]	T
Teeth	0.0373	0.0048	2.5491e-007	[iy]	T
Be	0.0441	1.3690	5.4635	[uw]	F

Table 5.2 Vowel [iy] classification results (Test words from dialect region 6, 7, and 8)

Test Word	[iy]	[ae]	[uw]	Results	True/False
Me	3.4432	1.6135	0.0110	[iy]	T
She	2.6936	0.5370	0.0011	[iy]	T
Me	1.4594	0.0299	1.6477e-005	[iy]	T
Seed	1.9629	4.0600	0.1304	[ae]	F
Meet	2.2924	0.0954	1.0848e-004	[iy]	T
Me	1.4456	0.1009	1.0457e-004	[iy]	T
She	3.4418	0.0921	9.3887e-005	[iy]	T
Me	3.4032	0.8794	0.0031	[iy]	T
Meet	0.7894	0.0067	1.4790e-006	[iy]	T
Meat	1.5498	2.8478	0.0649	[ae]	F
Feel	0.0181	0.9070	4.9601	[uw]	F
Be	2.4044	0.3258	7.8987e-004	[iy]	T
Free	3.2864	1.4014	0.0078	[iy]	T
Need	3.1481	0.0500	4.5818e-005	[iy]	T
We	0.6757	0.1017	2.7374e-005	[iy]	T
Sleep	1.0601	0.3789	0.0010	[iy]	T

Table 5.3 Vowel [ae] classification results (Test words from dialect region 1 and 2)

Test Word	[iy]	[ae]	[uw]	Results	True/False
Had	1.9017	4.4157	0.1641	[ae]	T
Rag	1.7942	4.7218	0.2017	[ae]	T
That	0.5511	4.4772	0.6107	[ae]	T
Had	2.1412	0.0179	7.4690e-006	[iy]	F
Had	0.1084	0.7910	0.1014	[ae]	T
Rag	0.8559	5.1009	0.6537	[ae]	T
That	0.2835	3.9539	1.1760	[ae]	T
That	1.1301e-004	0.0063	3.5959	[uw]	F
At	3.1258	1.0221	0.0059	[iy]	F
Cash	0.1767	3.0246	2.8895	[ae]	T
Can	0.0299	1.2981	1.8106	[uw]	F
Hat	0.0013	0.0886	6.4182	[uw]	F
Rag	0.0379	0.1859	6.0846e-004	[ae]	T
That	1.2648	4.2922	0.2192	[ae]	T

Table 5.4 Vowel [ae] classification results (Test words from dialect region 3, 4, 5 6, 7, and 8)

Test Word	[iy]	[ae]	[uw]	Results	True/False
Had	0.6167	4.8773	0.9979	[ae]	T
Rag	0.4860	5.0196	1.4589	[ae]	T
That	2.1270	4.1915	0.1267	[ae]	T
back	0.3792	0.2169	1.1689e-004	[iy]	F
Black	2.1048	4.2813	0.1374	[ae]	T
Lack	0.0255	0.9432	6.2033	[uw]	F
Panic	0.3087	0.4068	0.0015	[ae]	T
Had	0.5011	4.4744	1.1474	[ae]	T
Stag	1.7019	1.5360	0.0092	[iy]	F
Ask	1.0797	5.4741	0.5392	[ae]	T
Rag	3.2738	2.0782	0.0192	[iy]	F
That	0.0175	0.9185	1.3530	[ae]	T
Ask	1.3750	1.3254	0.0122	[iy]	F
Rag	3.7053	0.2440	4.9355e-004	[iy]	F
Had	0.1439	3.1702	2.7415	[ae]	T
That	1.4847	5.0318	0.2904	[ae]	T
Sat	1.8678	2.8495	0.0585	[ae]	T
Had	1.5456	3.8510	0.1257	[ae]	T
That	0.2328	3.7946	2.6643	[ae]	T
Rag	3.5680	0.6624	0.0019	[iy]	F
Rag	1.0565e-004	0.0162	1.6558	[uw]	F
That	0.7790	5.2448	0.8151	[ae]	T
Ask	0.8900	2.6792	0.0792	[ae]	T
Rag	0.0021	0.2183	1.1383	[uw]	F
Tax	0.1698	3.0576	3.0982	[uw]	F
Cat	1.3982	4.2371	0.1867	[ae]	T
Ask	1.3916	0.5453	0.0020	[iy]	F
Rag	0.5780	1.6931	0.0324	[ae]	T
That	0.1998	0.9577	0.0151	[ae]	T
Lamp	1.1234	1.3247	0.0130	[ae]	T
Has	0.1270	2.9389	2.1304	[ae]	T
Had	0.6908	0.1866	8.4425e-005	[iy]	F
Can	1.5629	0.1897	1.2654e-004	[iy]	F
Lad	0.1261	0.9953	0.1666	[ae]	T
Rag	0.4930	0.6468	0.0035	[ae]	T

Table 5.5 Vowel [uw] classification results (Test words from dialect region 1, 2, 3, 4, 5, 6, 7, and 8)

Test Word	[iy]	[ae]	[uw]	Results	True/False
Foot	0.8213	5.4862	0.7896	[ae]	F
Moon	0.0879	1.5569	2.3259	[uw]	T
Fool	1.2408e-006	1.2283e-004	0.7195	[uw]	T
Soon	3.7926	0.4643	0.0011	[iy]	F
Room	8.7932e-005	0.0114	2.3970	[uw]	T
Woods	1.7605e-004	0.0105	4.2055	[uw]	T
Zoo	3.2690e-004	0.0064	0.7443	[uw]	T
Zoo	0.0102	0.4198	6.5380	[uw]	T
Noon	0.9355	0.5336	8.1590e-004	[iy]	F
Noon	2.4200e-005	0.0057	0.3705	[uw]	T
Roof	0.0890	0.3894	0.0024	[ae]	F
Roof	8.5537e-004	0.0329	1.8255e-004	[ae]	F
Roof	0.0018	0.0578	3.1300	[uw]	T
Choose	0.9953	5.4777	0.6187	[ae]	F
Noon	0.2748	2.4764	0.1924	[ae]	F
Noon	0.8611	0.0979	2.8725	[uw]	T
Toon	0.0098	0.5410	6.2739	[uw]	T
Too	3.5667e-004	0.0070	0.7549	[uw]	T
Room	3.5359	0.1786	3.0455e-004	[iy]	F
Choose	9.7114e-005	0.1303	2.3244	[uw]	T
Rooms	2.1913	1.4492	0.0127	[iy]	F
Zoo	0.0286	1.2796	2.3624	[uw]	T
Too	2.8904	0.8119	0.0024	[iy]	F
Soon	2.7643	2.7450	0.0427	[iy]	F
Choose	3.9401	1.0193	0.0050	[iy]	F
Room	0.0432	0.1866	1.1122	[uw]	T
Room	0.0054	0.2201	0.3680	[uw]	T
Room	7.3359e-004	0.0908	1.9293	[uw]	T
Tooth	0.0603	1.9774	3.9577	[uw]	T
Sooth	0.0704	1.7446	0.5142	[ae]	F
Roof	0.0045	0.3626	3.1093	[uw]	T
Noon	0.0091	1.5120	3.1730	[uw]	T
Proof	0.9861	4.8606	0.3940	[ae]	F
Tooth	0.2971	2.4763	0.5318	[ae]	F
Zoo	0.0058	0.3019	7.0480	[uw]	T
Zoo	0.0074	0.5061	0.7032	[uw]	T
Roof	0.0587	0.4597	0.0054	[ae]	F
Proof	3.1243e-004	0.0497	0.7984	[uw]	T
Proof	0.0124	0.0238	0.4295	[uw]	T

Table 5.6 Consonant [sh] classification results (Test words from dialect region 1,2, 3, 4, 5, 6, 7, and 8)

Test Word	[sh]	[f]	Results	True/False
She	2.4826	2.4954	[f]	F
Wash	0.9184	7.7010	[f]	F
Wash	0.0086	0.0014	[sh]	T
She	3.0659	0.1465	[sh]	T
Wash	1.6108	0.9333	[sh]	T
She	5.3151	0.0302	[sh]	T
Wash	5.1490	0.1073	[sh]	T
Wash	4.3749	0.0789	[sh]	T
She	4.7404	0.0196	[sh]	T
She	0.0031	0.0703	[f]	F
Fish	3.7395	0.2649	[sh]	T
Fish	5.1277	0.0391	[sh]	T
Wash	2.1695	2.5148	[f]	F
She	2.1008	8.7738e-004	[sh]	T
Wash	4.3170	0.2299	[sh]	T
She	1.9581	3.0584e-004	[sh]	T
Shot	0.7064	2.1771e-004	[sh]	T
She	2.3077	0.0897	[sh]	T
Should	4.8045	0.0056	[sh]	T
Wash	4.9615	0.1437	[sh]	T
She	1.5637	0.06317	[sh]	T
Should	3.1919	7.8590e-004	[sh]	T
Fresh	5.3819	0.0348	[sh]	T
Dish	2.3352	9.5467e-005	[sh]	T
Wash	4.8205	0.0056	[sh]	T
She	0.7884	3.4866e-006	[sh]	T
Should	5.3314	0.0329	[sh]	T
Shelt	0.0103	9.8455e-007	[sh]	T
She	3.5569	1.3674	[sh]	T
She	3.2165	9.5528e-004	[sh]	T
Wash	3.2778	4.1324e-0047	[sh]	T
She	1.6542	1.8567e-0057	[sh]	T
Wash	3.2728	0.0018	[sh]	T
Brush	4.4522	0.1617	[sh]	T
She	3.2910	8.6683e-004	[sh]	T
Shut	5.1188	0.0544	[sh]	T
Shrimp	2.5228	0.0028	[sh]	T
Wash	3.8399	0.1958	[sh]	T
Wash	3.5374	1.3949	[sh]	T
Ship	5.3622	0.0285	[sh]	T

Table 5.7 Consonant [f] classification results (Test words from dialect region 1,2, 3, 4, 5, 6, 7, and 8)

Test Word	[sh]	[f]	Results	True/False
Far	1.2327	6.2735	[f]	T
Fool	0.0422	1.0856	[f]	T
Off	0.0374	5.7884	[f]	T
For	0.2558	9.1685	[f]	T
Foot	0.4066	10.0661	[f]	T
Fish	0.2799	9.6487	[f]	T
Feel	1.1337	6.8770	[f]	T
Leaf	0.9017	7.1849	[f]	T
Roof	1.3641	3.7811	[f]	T
Roof	0.0141	4.7026	[f]	T
Enough	2.0005	3.8533	[f]	T
Roof	0.0588	2.6795	[f]	T
Feed	9.7096e-004	0.6403	[f]	T
From	0.7740	8.2306	[f]	T
If	1.1843	5.2007	[f]	T
If	0.3314	10.2137	[f]	T
Enough	1.1275	6.6849	[f]	T
Roof	0.0574	8.0399	[f]	T
Fresh	0.78417	5.5536	[f]	T
Flee	0.3665	9.4438	[f]	T
First	3.1265e-004	0.4041	[f]	T
Enough	0.0129	1.1430	[f]	T
Of	0.1800	5.3871	[f]	T
For	0.1613	10.3838	[f]	T
First	0.6551	3.7237	[f]	T
Jeff	0.1878	4.3599	[f]	T
Fish	0.2284	2.6212	[f]	T
Enough	0.9598	4.7870	[f]	T
Form	0.6903	5.2687	[f]	T
For	0.0483	3.3340	[f]	T
Roof	0.1941	8.9762	[f]	T
Proof	0.1331	9.4462	[f]	T
If	0.0085	3.1391	[f]	T
Free	0.2439	0.15257	[sh]	F
Proof	0.0518	5.1547	[f]	T
Foam	0.0777	5.6398	[f]	T
Chief	0.0065	1.9503	[f]	T

Table 5.8 Summary of the vowel [iy] classification results in eight dialect regions

Dialect Region	Correct	Wrong	Total
DR1	6(100.00%)	0(0.00%)	6(12.00%)
DR2	6(54.55%)	5(45.45%)	11(22.00%)
DR3	5(83.33%)	1(16.67%)	6(12.00%)
DR4	3(60.00%)	2(40.00%)	5(10.00%)
DR5	4(66.67%)	2(33.33%)	6(12.00%)
DR6	4(80.00%)	1(20.00%)	5(10.40%)
DR7	4(66.67%)	2(33.33%)	6(12.00%)
DR8	5(100.00%)	0(0.00%)	5(10.00%)

Table 5.9 Summary of the vowel [ae] classification results in eight dialect regions

Dialect Region	Correct	Wrong	Total
DR1	3(75.00%)	1(25.00%)	4(8.16%)
DR2	6(60.00%)	4(40.00%)	10(20.41%)
DR3	3(75.00%)	1(25.00%)	4(8.16%)
DR4	5(62.50%)	3(37.50%)	8(16.33%)
DR5	2(50.00%)	2(50.00%)	4(8.16%)
DR6	6(60.00%)	4(40.00%)	10(20.41%)
DR7	4(66.67%)	2(33.33%)	6(12.24%)
DR8	2(66.67%)	1(33.33%)	3(6.12%)

Table 5.10 Summary of the vowel [uw] classification results in eight dialect regions

Dialect Region	Correct	Wrong	Total
DR1	3(60.00%)	2(40.00%)	5(12.82%)
DR2	4(57.14%)	3(42.86%)	7(17.95%)
DR3	3(60.00%)	2(40.00%)	5(12.82%)
DR4	3(60.00%)	2(40.00%)	5(12.82%)
DR5	3(50.00%)	3(50.00%)	6(15.38%)
DR6	3(60.00%)	2(40.00%)	5(12.82%)
DR7	3(60.00%)	2(40.00%)	5(12.82%)
DR8	1(100.00%)	0(0.00%)	1(2.56%)

Table 5.11 Summary of vowels classification results

Vowel	Correct	Wrong	Total
[iy]	37(74.00%)	13(26.00%)	50(36.23%)
[ae]	31(63.27%)	18(36.73%)	49(35.51%)
[uw]	23(58.97%)	16 (41.03%)	39(28.26%)
Total	90(65.22%)	48(34.78%)	138(100.00%)

Table 5.12 Summary of the consonant [sh] classification results in eight dialect regions

Dialect Region	Correct	Wrong	Total
DR1	6(75.00%)	2(25.00%)	8(20.00%)
DR2	3(60.00%)	2(40.00%)	5(12.50%)
DR3	7(100.00%)	0(0.00%)	7(17.50%)
DR4	3(100.00%)	0(0.00%)	3(7.50%)
DR5	5(100.00%)	0(0.00%)	5(12.50%)
DR6	5(100.00%)	0(0.00%)	5(12.50%)
DR7	3(100.00%)	0(0.00%)	3(7.50%)
DR8	4(100.00%)	0(0.00%)	4(10.00%)

Table 5.13 Summary of the consonant [f] classification results in eight dialect regions

Dialect Region	Correct	Wrong	Total
DR1	5(100.00%)	0(0.00%)	5(13.51%)
DR2	6(100.00%)	0(0.00%)	6(16.22%)
DR3	5(100.00%)	0(0.00%)	5(13.51%)
DR4	3(100.00%)	0(0.00%)	3(8.11%)
DR5	5(100.00%)	0(0.00%)	5(13.51%)
DR6	5(100.00%)	0(0.00%)	5(13.51%)
DR7	3(100.00%)	0(0.00%)	3(8.11%)
DR8	4(80.00%)	1(20.00%)	5(13.51%)

Table 5.14 Summary of consonants classification results

Consonant	Correct	Wrong	Total
[sh]	36(90.00%)	4(10.00%)	40(51.95%)
[f]	36(97.30%)	1(2.70%)	37(48.05%)
Total	72(93.51%)	5(6.49%)	77(100.00%)

Chapter 6

CONCLUSIONS AND FURTHER STUDY

6.1 Conclusions

Linear predictive analysis is one of the most powerful speech analysis techniques. Since the method can estimate the basic speech parameters, e.g., pitch, formants, spectra and vocal tract area functions, and since it represents speech for low bit rate transmission or storage, it has become the predominant technique. The ability to provide extremely accurate estimates of the speech parameters and the relative speed of computation decide the importance of the linear predictive method in speech analysis area.

One of the most important applications of the linear predictive analysis in speech is the area of low bit rate encoding of speech for transmission (LPC vocoder). In the LPC vocoder system, the speech synthesis model shows that speech can be modelled as the output of a linear, successive time-varying system excited by either quasi-periodic pulses (for voiced sound) or random noise (for unvoiced sound). In order to characterize the linear, time-varying system, it can be realized by linear, successive and time-invariant filters. The linear prediction method can estimate the parameters which characterize the successive and time-invariant filters. In fact, the linear predictive method determines robust and reliable parameters associated with an all-pole IIR filter in speech synthesis. The IIR filter can be realized by PARCOR parameters in a lattice structure filter.

Speech recognition by machine (in many contexts also known as automatic speech recognition, computer speech recognition) is the process of converting a speech signal to a set of words or interpretation of everything the human speaker

says while the machine is listening. To recognizing the basic phonemes, syllables and isolated words in continuous speech is the fundamental purpose of speech recognition systems. In this thesis, we make use of PARCOR parameters in LPC vocoder systems to realize the recognition of phonemes in a continuous speech.

In Chapter 4, we present the distributions of PARCOR parameters of vowels and consonants in Figure 4.11 and Figure 4.12. The data in both figures are spoken by female and male people who are from from eight dialect regions. In Figure 4.11 (a) and (b), we can see the cluster of each vowel [iy], [ae] and [uw], which are marked by star, circle and plus, are well separated. Also in the Figure 4.12 (a) and (b), it is obvious that we can find the partition line between the cluster of [sh] and [f]. They indicate that the PARCOR parameters have the potential capability to characterize phonemes.

In Chapter 5, we explore a method to classify the phoneme in one-syllable word by using a supervised classifier. The supervised classifier need to be trained. The training uses TIMIT speech database, which contains the recordings of 630 speakers of 8 major dialects of American English. The training data were grouped into the vowel group including phoneme [ae], [iy] and [uw] and the consonant group including [sh] and [f]. After the training, the decision rule was derived. Then we designed two classifiers to classify the unknown phonemes in one-syllable words. Before classification, the test words needed to be preprocessed to detect the phonemes and calculate the corresponding PARCOR parameters. In Figure 5.9, the preprocessing method is illustrated. By using this method, the frames can be grouped into vowel, consonant and silence group. Figure 5.13 shows the group results of vowel, consonant phonemes and silence in the word "cat" and "she". It indicates the preprocessing method is feasible to detect phonemes and group them into vowel, consonant or silence. The calculated vowel and consonant PARCOR parameters of frames are fed into the classifier, which uses the maximum likelihood decision rule to classify the unknown phonemes. The classification results are shown from Table 5.1 to Table 5.7. The correct classification rate, which is shown in Table 5.8 to Table 5.14 is good enough. In Table 5.11, we see the correct classification rate

of each vowel [iy], [ae] and [uw] were 74.00%, 63.27% and 58.97% respectively. The total correct classification rate was 65.22% for the vowel group. In Table 5.14, the correct classification rate of [sh] and [f] were 90.00% and 97.30% respectively. The total correct classification rate was 93.51% for the consonant group.

But in some cases, some phonemes can't be recognized correctly, partly it is because the variation of speech from different people who are from different dialect regions. For the training data, we found that the cluster of PARCOR parameters of each vowel in Figure 5.1 (b), was not completely separated, there is overlap between the cluster of [ae], [iy] and [uw]. In Figure 5.3 (b), the cluster of PARCOR parameters of [sh] and [f] is not completely separated, either. That's why some phonemes can't be classified correctly. Overall, the results of classification are good enough. They indicate PARCOR parameters have potential ability to characterize the phonemes.

6.2 Further Study

In this thesis, we investigate a few vowel and consonant phonemes, more vowel and consonant phonemes are should be studied and syllables should be classified in the future.

For syllables, they are typically made up of a syllable nucleus (most often a vowel) with optional initial and final margins (typically, consonants), first stage, we classify vowel or consonant phonemes by using the vowel or consonant classifier, then combine the classification results from both classifiers, finally judge the unknown syllable belong to which class. For example, one unknown one-syllable word, after acquiring the corresponding PARCOR parameters of vowel and consonant frame by using the method shown in Figure 5.9, we feed vowel and consonant PARCOR parameters into the vowel and consonant classifier respectively, if the vowel classifier output is [iy] and the the consonant clarifier output is [sh], plus considering the sequence of the frame occurred in the one-syllable word, we can judge the syllable is "she".

Also how the combination of consonants and vowels affect the sound should be

further studied, for example, the vowel [uw] in the one-syllable word "wood" often is spoken as [uh].

Bibliography

[1] Lawrence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[2] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 Kbits/s," *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1466–1474, 1975.

[3] Blade Kotelly, *The Art and Business of Speech Recognition*, Pearson Education, Inc., Boston, MA, 2003.

[4] "Dragon naturallyspeaking," http://www.dragonsys.ca//.

[5] Albert S. Bregman, *Auditory Scene Analysis*, MIT Press., Cambridge, Mass., 1990.

[6] Martin P. Cooke, Andrew C. Morris and Philip D. Green, "Missing data techniques for robust speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, April 1997, vol. 2, pp. 863–866.

[7] Shigeru Katagiri, Ed., *Handbook of Neural Networks for Speech Processing*, pp. 63–117, Artech House, Boston.London, 2000.

[8] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.

[9] David P. Morgan and Christopher L. Scofield, *Neural Networks and Speech Processing*, Kluwer Academic Publishers, Boston/Dordrecht/London, 1991.

[10] Lawrence R. Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1993.

[11] "A beginner's guide to phonetics," http://www.jcarreras.homestead.com/RRPhonetics1.html English Language & Linguistics Programme, Introductory Phonetics at Roehampton University.

[12] Shigeru Katagiri, Ed., *Handbook of Neural Networks for Speech Processing*, pp. 19–62, Artech House, Boston.London, 2000.

[13] National Institute of Standards and Technology (NIST), *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, NIST Speech Disc CD1-1.1 edition, October 1990.

[14] Harold T. Edwards, *Applied Phonetics the Sounds of American English*, Singular Publishing Group, Inc., San Diego.London, 1997.

[15] Lajos Hanzo, F. Clare A. Somerville and Jason P. Woodard, *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels*, John Wiley & Sons, New York, 2001.

[16] N. Jayant and P. Noll, *Digital Coding of Waveforms, Principles and Applications to Speech and Video*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1984.

[17] A. M. Kondoz, *Digital Speech*, John Wiley & Sons, New York, 1994.

[18] John Makhoul, "Linear prediction: a tutorial review," in *Proceedings of the IEEE*, April 1975, vol. 63, pp. 561–580.

[19] Simon Haykin, "Radar sinal processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing(ASSP)*, vol. 2, no. 2, pp. 2–18, April 1985.

[20] John D. Markel and Augustine H. Gray, *Linear Prediction of Speech*, Springer-verlag, New York, 1976.

[21] Charles W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1992.

[22] Boaz Porat, *A Course in Digital Signal Processing*, John Wiley & Sons, New York, 1997.

[23] Enders A. Robinson and Sven Treitel, "Maximum entropy and the relationship of the partial autocorrelation to the reflection coefficients of a layered system," *IEEE Transactions on Acoustics, Speech, and Signal Processing(ASSP)*, vol. 28, no. 2, pp. 224–235, April 1980.

[24] Akihiro Taguchi and Kunio Takaya, "Capability of classifying vowels with a residual excited linear predication (RELP) vocoder," in *2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing PACRIM'03*, Victoria, Canada, August 2003, pp. 310–313.

[25] Ying Cui and Kunio Takaya, "Recognition syllables in a continuous stream of speech by PARCOR parameters of linear predictive vocoder," in *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, Saskatoon, Canada, May 2005, pp. 2271–2274.

[26] Akihiro Taguchi, "Residual-excited linear predictive (RELP) vocoder system with TMS320C6711 DSK and vowel characterization," M.S. thesis, Department of Electrical Engineering, University of Saskatchewan, Saskatoon, December 2003.

[27] William S. Meisel, *Computer-Oriented Approaches to Pattern Recognition*, Academic Press, New York, 1972.

Appendix A

TIMIT CORPUS

TIMIT was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The corpus was prepared at the National Institute of Standards and Technology (NIST) with sponsorship from the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO).

Table A.1 Dialect distribution of speakers

Dialect Region (DR)	**Male**	**Female**	**Total**
DR1: New England	31(63%)	18 (27%)	49 (8%)
DR2: Northern	71 (70%)	31 (30%)	102 (16%)
DR3: North Midland	79 (67%)	23 (23%)	102 (16%)
DR4: South Midland	69 (69%)	31 (31%)	100 (16%)
DR5: Southern	62 (63%)	36 (37%)	98 (16%)
DR6: New York City	30 (65%)	16 (35%)	46 (7%)
DR7: Western	74 (74%)	26 (26%)	100 (16%)
DR8: Army Brat (moved around)	22 (67%)	11 (33%)	33 (5%)
Total eight DR	438 (70%)	192 (30%)	630 (100%)

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. Table A.1 shows the number of speakers for the 8 dialect regions, broken down by sex. The percentages are given in parentheses. A speaker's dialect region is the geographical area of the U.S. where they lived during their childhood years. The geographical areas correspond with recognized dialect regions in U.S. (Language Files, Ohio State University Linguistics Dept., 1982), with the exception of the Western region (DR7)

in which dialect boundaries are not known with any confidence and dialect region 8 where the speakers moved around a lot during their childhood. All the phonemic and phonetic symbols used in the TIMIT lexicon and in the phonetic transcriptions are shown in Table A.2

Table A.2 Phonemic and phonetic symbols from TIMIT speech corpus

Symbol	Example	Symbol	Example
iy	beet	m	mom
ih	bit	n	noon
eh	bet	ng	sing
ey	bay	em	bottom
ae	cat	en	button
aa	bott	eng	washington
aw	bout	nx	winner
ay	bike	b	bee
ah	but	d	dog
ao	bought	g	get
oy	toy	p	put
ow	boat	t	too
uh	book	k	kite
uw	boot	dx	dirty
ux	toot	q	bat
ax	about	jh	joke
ix	debit	ch	choke
axr	butter	s	sea
ax-h	suspect	sh	sheep
er	bird	dh	then
l	lay	z	zone
r	ray	zh	azure
w	way	f	fun
y	yacht	th	thin
hh	hay	v	van
hv	ahead	pau	paues
el	bottle	epi	epenthetic silence

Appendix B

TEST DATA

The information of the test words, such as speaker name, gender and dialect region, are listed in Table B.1 to Table B.7.

Table B.1 Test words for vowel [iy] from dialect region 1, 2, 3, 4 and 5

Test Word	Region	Speaker Name	Gender
She	DR1	fadw0	Female
She	DR1	fdml0	Female
Me	DR1	faks0	Female
Seem	DR1	faks0	Female
Be	DR1	mdab0	Male
Seed	DR1	mjsw0	Male
She	DR2	fjas0	Female
She	DR2	fcmr0	Female
He	DR2	fdrd1	Female
Clean	DR2	fdrd1	Female
Street	DR2	mabw0	Male
Feel	DR2	mabw	Male
Leaf	DR2	mrfk0	Male
Each	DR2	fjre0	Female
Real	DR2	fjre0	Female
Cream	DR2	mccs0	Male
Me	DR2	mccs0	Male
Feed	DR3	fcmh0	Female
She	DR3	fcmh0	Female
Me	DR3	fcmh0	Female
Tea	DR3	fcmh0	Female
Me	DR3	mbdg0	Male
Steep	DR4	fadg0	Female
Read	DR4	fcft0	Female
She	DR4	fdms0	Female
Me	DR4	mjrf0	Male
Be	DR4	mjrf0	Male
Beach	DR4	mkcl0	Male
Flee	DR5	fasw0	Female
Seem	DR5	fawf0	Female
See	DR5	fgmd0	Female
Cheap	DR5	mahh0	Male
Teeth	DR5	mahh0	Male
Be	DR5	fcal1	Female

Table B.2 Test words for vowel [iy] from dialect region 6, 7, and 8

Test Word	Region	Speaker Name	Gender
Me	DR6	fdrw0	Female
She	DR6	fdrw0	Female
Me	DR6	fmgd0	Female
Seed	DR6	mesd0	Male
Meet	DR6	flnh0	Female
Me	DR7	fisb0	Female
She	DR7	fisb0	Female
Me	DR7	fdhc0	Female
Meet	DR7	fisb0	Female
Meat	DR7	mgrt0	Male
Feel	DR7	mgrt0	Male
Be	DR8	fcmh1	Female
Free	DR8	fcmh1	Female
Need	DR8	fjsj0	Female
We	DR8	mjln0	Male
Sleep	DR8	mdaw1	Male

Table B.3 Test words for vowel [ae] from dialect region 1 and 2

Test Word	Region	Speaker Name	Gender
Had	DR1	fdaw0	Female
Rag	DR1	faks0	Female
That	DR1	faks0	Female
Had	DR1	mreb0	Male
Had	DR2	fjre0	Female
Rag	DR2	fjre0	Female
That	DR2	fjre0	Female
That	DR2	fcmr0	Female
At	DR2	fjwb0	Female
Cash	DR2	fjwb0	Female
Can	DR2	mabw0	Male
Hat	DR2	mdlb0	Male
Rag	DR2	mcss0	Male
That	DR2	mcss0	Male

Table B.4 Test words for vowel [ae] from dialect region 3, 4, 5, 6, 7 and 8

Test Word	Region	Speaker Name	Gender
Had	DR3	fcmh0	Female
Rag	DR3	fjlr0	Female
That	DR3	mddc0	Male
back	DR3	mdwm0	Male
Black	DR4	fcrh0	Female
Lack	DR4	fadg0	Female
Panic	DR4	fdms0	Female
Had	DR4	fdms0	Female
Stag	DR4	mcdr0	Male
Ask	DR4	mjrf0	Male
Rag	DR4	mjrf0	Male
That	DR4	mjrf0	Male
Ask	DR5	fcal1	Female
Rag	DR5	fcal1	Female
Had	DR5	fawf0	Female
That	DR5	mcrc0	Male
Sat	DR5	mctt0	Male
Had	DR6	fdrw0	Female
That	DR6	fdrw0	Female
Rag	DR6	fdrw0	Female
Rag	DR6	fmgd0	Female
That	DR6	fmgd0	Female
Ask	DR6	fmgd0	Female
Rag	DR6	mdsc0	Male
Tax	DR6	flnh0	Female
Cat	DR6	mjdh0	Male
Ask	DR7	fisb0	Female
Rag	DR7	fisb0	Female
That	DR7	fisb0	Female
Lamp	DR7	mers0	Male
Has	DR7	mgrt0	Male
Had	DR7	mdlf0	Male
Can	DR8	fcmh1	Female
Lad	DR8	fcmh1	Female
Rag	DR8	mdaw1	Male

Table B.5 Test words for vowel [uw] from dialect region 1,2, 3, 4, 5, 6, 7 and 8

Test Word	Region	Speaker Name	Gender
Foot	DR1	mstk0	Male
Moon	DR1	mwar0	Male
Fool	DR1	faks0	Female
Soon	DR1	mrjo0	Male
Room	DR1	msjs1	Male
Woods	DR2	fdrd1	Female
Zoo	DR2	feac0	Female
Zoo	DR2	mrhl0	Male
Noon	DR2	mrjm0	Male
Noon	DR2	mrjt0	Male
Roof	DR2	mjae0	Male
Roof	DR2	mdlc2	Male
Roof	DR3	fmjf0	Female
Choose	DR3	fcmg0	Female
Noon	DR3	fdfb0	Female
Noon	DR3	mcal0	Male
Toon	DR3	mcal0	Male
Too	DR4	fdms0	Female
Room	DR4	mjrf0	Male
Choose	DR4	mesg0	Male
Rooms	DR4	mkcl0	Male
Zoo	DR4	mfrm0	Male
Too	DR5	fasw0	Female
Soon	DR5	fawf0	Female
Choose	DR5	fmpg0	Female
Room	DR5	fhes0	Female
Room	DR5	fjsa0	Female
Room	DR5	fmah0	Female
Tooth	DR6	mesd0	Male
Sooth	DR6	mjdh0	Male
Roof	DR6	majp0	Male
Noon	DR6	fklc1	Female
Proof	DR6	fmju0	Female
Tooth	DR7	fgwr0	Female
Zoo	DR7	mjfr0	Male
Zoo	DR7	mtml0	Male
Roof	DR7	mgsl0	Male
Proof	DR7	mcth0	Male
Proof	DR8	mpam0	Male

Table B.6 Test words for consonant [sh] from dialect region 1,2, 3, 4, 5, 6, 7 and 8

Test Word	Region	Speaker Name	Gender
She	DR1	fdaw0	Female
Wash	DR1	fdaw0	Female
Wash	DR1	fdml0	Female
She	DR1	fjem0	Female
Wash	DR1	fjem0	Female
She	DR1	mwbt0	Male
Wash	DR1	mwbt0	Male
Wash	DR1	mstk0	Male
She	DR2	fjas0	Female
She	DR2	fcmr0	Female
Fish	DR2	fdrd1	Female
Fish	DR2	mwvw	Male
Wash	DR2	mwvw	Male
She	DR3	fcmh0	Female
Wash	DR3	fpkt0	Female
She	DR3	fpkt0	Female
Shot	DR3	mdwm0	Male
She	DR3	mthc0	Male
Should	DR3	mthc0	Male
Wash	DR3	mthc0	Male
She	DR4	fdms0	Female
Should	DR4	mrgm0	Male
Fresh	DR4	mbns0	Male
Dish	DR5	futb0	Female
Wash	DR5	fnlp0	Female
She	DR5	msfh1	Male
Should	DR5	mrws1	Male
Shelt	DR5	mrrk0	Male
She	DR6	fdrw0	Female
She	DR6	fmgd0	Female
Wash	DR6	fmgd0	Female
She	DR6	mrjr0	Male
Wash	DR6	mrjr07	Male
Brush	DR7	fgwr0	Female
She	DR7	fisb0	Female
Shut	DR7	mdvc0	Male
Shrimp	DR8	fmld0	Female
Wash	DR8	mres0	Male
Wash	DR8	mpam0	Male
Ship	DR8	majc0	Male

Table B.7 Test words for consonant [f] from dialect region 1,2, 3, 4, 5, 6, 7 and 8

Test Word	Region	Speaker Name	Gender
Far	DR1	faks0	Female
Fool	DR1	faks0	Female
Off	DR1	fjem0	Female
For	DR1	mwbt0	Male
Foot	DR1	mstk0	Male
Fish	DR2	fdrd1	Female
Feel	DR2	mabw0	Male
Leaf	DR2	mrfk0	Male
Roof	DR2	mjae0	Male
Roof	DR2	mdlc2	Male
Enough	DR2	mabw0	Male
Roof	DR3	fmjf0	Female
Feed	DR3	fcmh0	Female
From	DR3	fpkt0	Female
If	DR3	mwjg0	Male
If	DR3	mtdt0	Male
Enough	DR4	fdms0	Female
Roof	DR4	flhd0	Female
Fresh	DR4	mbns0	Male
Flee	DR5	fasw0	Female
First	DR5	futb0	Female
Enough	DR5	fnlp0	Female
Of	DR5	mrws1	Male
For	DR5	mrrk0	Male
First	DR6	fdrw0	Female
Jeff	DR6	mcmj0	Male
Fish	DR6	mcmj0	Male
Enough	DR6	mdsc0	Male
Form	DR6	mjfc0	Male
For	DR7	fcau0	Female
Roof	DR7	mgsl0	Male
Proof	DR7	mcth0	Male
If	DR8	fjsj0	Female
Free	DR8	fcmh1	Female
Proof	DR8	mpam0	Male
Foam	DR8	mdaw1	Male
Chief	DR8	mjln0	Male